Production training and contextual similarity hurt the comprehension of new vocabulary

Megan Waller[1], Daniel Yurovsky[1] & Nazbanou Nozari[2,3]

[1] Department of Psychology, Carnegie Mellon University

[2] Department of Psychological and Brain Sciences, Indiana University

[3] Cognitive Science Program, Indiana University

Correspondence concerning this article should be addressed to Megan Waller.

Department of Psychology, Carnegie Mellon University

5000 Forbes Avenue, Pittsburgh, PA, 15213, United States

meganwal@andrew.cmu.edu

Production training and contextual similarity hurt the comprehension of new vocabulary

Abstract

In two experiments (N = 179), we studied the effects of contextual similarity and training mode on the comprehension of new vocabulary. Participants were trained on new vocabulary in blocks of semantically-similar, phonologically-similar, or unrelated items. Each participant was trained through passive exposure, active comprehension, or active production. Same number of items were trained in clusters of 9 in Experiment 1 and clusters of 3 in Experiment 2, manipulating difficulty during training. Results showed a detrimental and persistent effect of semantic similarity, and a less robust effect of phonological similarity, both of which grew larger over time. We also found a negative and largely independent influence of production mode on learning, which, contrary to the similarity effect, shrank with time. Neither effect was modulated by difficulty at training time. These findings shed further light on the factors influencing new vocabulary learning and open new avenues for larger-scale and classroom-level studies.

Keywords: vocabulary learning; word production; contextual similarity; semantic interference; phonological interference

Introduction

Although learning a new language has many facets, learning the relationship between words and the meaning they specify, i.e., vocabulary learning, is a basic block in the process. This is true for children learning their first language, for neurotypical adults learning a new language, and even for individuals with brain damage such as stroke survivors re-learning their lost language. Naturally, a critical question is "What is the most efficient way to teach/learn new words?". This paper tackles this issue in learning a new language in neurotypical adult participants. A large and diverse literature has investigated this issue, and has uncovered various principles for improving learning. Some of these principles remain the subject of dispute. For instance, there is an ongoing debate on whether errorless learning (i.e., preventing learners from making mistakes by modeling the utterance for them) or errorful learning (i.e., allowing learners to make mistakes and then correcting them) is the optimal way to learn vocabulary (Fillingham et al. 2006; Middleton et al., 2015; Saxton, 1997; Saxton et al., 1998; Shuchard & Middleton, 2018; Waller et al., 2024). Other principles are more widely accepted. For example, learners generally benefit from spacing (interleaving training items) over massed practice (repeated studying of the same item; Kornell, 2009; Nakata & Suzuki, 2019; see Kim & Webb, 2022 for a meta-analysis), and from deeper processing of the materials during study (Craik & Tulving, 1975; Jacoby et al., 2005; Lawrence et al., 2024).

This paper tackles two factors, whose effect on vocabulary learning remains controversial: *learning mode* and *contextual similarity*. Learning mode refers to the training method and its interaction with learning goals. For example, if the goal is for learners to be able to comprehend new words, is it better to train them using a comprehension or a production training mode? Contextual similarity refers to the relationship between items in a training set. Most pedagogical

settings, as well as language learning apps, such as Babble (https://www.babbel.com/), Duolingo (https://www.duolingo.com/), and Memrise (https://www.memrise.com/en-us/), group new words into semantically related categories, e.g., “animals”, “clothing items”, “fruits”, etc. But does semantic similarity facilitate or hinder learning? How about phonological similarity? Is it easier or harder to learn similar-sounding words such as “cap”, “map”, “cat”, “mat” together in one set?

Both factors have been the subject of many past studies, and the results are very mixed. While some studies have reported a benefit of production training for improving comprehension and perception (e.g. Bixby, 2017; Hopman & MacDonald, 2018; Icht & Mama, 2015; MacLeod et al. 2010; Zamuner et al., 2016, see MacLeod & Bodner, 2017 for a review), other studies have reported the opposite (e.g. Baese-Berk, 2010, 2019; Baese-Berk & Samuel, 2016, 2022; Kapnoula & Samuel, 2022; Kaushanskaya & Yoo, 2011; Krashen, 2003; Leach & Samuel, 2007; VanPatten & Cadierno, 1993; VanPatten, 2013; Zamuner et al., 2018). Similarly, while some studies have reported positive effects of contextual similarity on learning (e.g. Feng et al., 2022; Grandy, 2012; Hashemi & Gowdasiaei, 2005; Haycraft, 1978; Hoshino, 2010; Schneider et al., 2002; Seal, 1991; Stoller & Grabe, 1993; Wharton & Race, 1999), others have reported the opposite (e.g. Breining et al., 2019; Finkbeiner & Nicol, 2003; Higa, 1963; Korochkina et al., 2021; Nation, 2000; Papagno & Vallar, 1992; Papathanasiou, 2009; Pérez-Serrano et al., 2022; Schneider et al., 1998; 2013; Tinkham, 1993, 1997; Waring, 1997; Wilcox & Medina, 2013).

These discrepancies could be, at least in part, due to different methodologies, populations, measures, and different loci of the language processing systems targeted in each study. For example, while Papagano and Vallar (1992) reported a detrimental effect of contextual similarity on learning for phonologically similar items, Hoshino et al.’s (2010) report of a facilitatory contextual effect was based on semantic relations. Similarly, while the detrimental effect of

production on learning in Baese-Berk and Samuel (2016)'s study was focused on learning new acoustic contrasts, Hopman and MacDonald's (2018)'s study was focused on learning new vocabulary and grammatical structures in sentence. Thus, perhaps, different principles govern facilitation and interference in different parts of the language processing system. While this is a theoretical possibility, empirical evidence suggests otherwise. For example, Breining et al. (2019) showed that both lexical learning (i.e., learning new mappings between words and concepts) and segmental learning (i.e., learning new mappings between words and their phonemes/letters) were subject to interference from both semantic and segmental similarity. Similarly, the negative effects of production on perception have been argued to result from attentional bottlenecks that could very well be general to any stage of processing (e.g., Baese-Berk & Samuel, 2022; Kapnoula and Samuel, 2022). In keeping with this, discrepancies have sometimes been observed even in studies targeting the same stage of processing (e.g., Hoshio et al., 2010 vs. Korochkina et al., 2021; Bixby, 2017 vs. Baese-Berk & Samuel, 2016; 2022). In what follows, we review two bodies of literature on training mode and contextual similarity in more depth. We conclude that the current state of the literature calls for more well-controlled empirical studies with enough statistical power to provide reliable effects. We then present two experiments which investigate the effect of training mode, contextual similarity, and their potential interaction on the acquisition of new vocabulary in neurotypical adults.

The perception-production link and learning

Most accounts of language processing agree that perception and production are closely linked (Denes & Pinson, 1993; Guenther, 2016; Hickok & Poeppel, 2007; Pickering & Garrod, 2013), albeit in a complicated way (Baese-Berk et al., 2024). This link is critical for both learning and monitoring language production (e.g., Bohland, Bullock, & Guenther, 2010; Hickok, 2012;

Levelt et al., 1999; Nozari, 2020). Usually, it is perception that is thought to affect production. For example, listeners cannot learn new words unless they first hear them. Similarly, perceptual monitoring entails hearing one's own production and subsequently making changes to it. Empirical evidence also shows the influence of perception on production (e.g., Houde & Jordan, 1998, 2002; Kim et al., 2011), which can be implicit and very rapid (e.g., Murphy et al., 2023). However, there is also a great deal of variability in the relationship between production and perception among learners (e.g., Bradlow et al., 1999; Wang et al., 2003), motivating a more controversial position, namely that production affects perception during learning.

In its most extreme form, the influence of production on perception is captured by theories of direct realism, which hold that the perception of speech sounds depends critically on speech motor gestures (Best, 1995; Fowler, 1986; Galantucci et al., 2006; Liberman et al., 1967; Liberman & Mattingly, 1985). But there are other, more general, reasons for why production could be advantageous for learning in the perception system. For one, the depth of processing account of memory (Craik & Tulving, 1975) suggests that deeper processing of the materials at the time of encoding results in better recall. One way in which deeper processing can be accomplished is by forcing the learner to retrieve the materials, instead of simply studying them. The so-called "testing effect" has empirical support in studies showing that learners who are tested after a study phase do better on a final test compared to those who simply repeat the study phase again (Bjork, 1988; Carpenter et al., 2006; Karpicke & Blunt, 2011; Roediger & Karpicke, 2006; Zaromb & Roediger, 2010). Additionally, overtly producing items requires learners to retrieve articulatory-motor plans associated with them that are not necessarily activated during silent study (Oppenheim & Dell, 2008), and this so-called "production-effect" has been reported to benefit long-term retention of information (the Production effect, Icht & Mama, 2015; MacLeod et al., 2010).

On the other hand, there are also theoretical reasons to posit that production may hurt perception and recognition. One idea is transfer-appropriate processing, the finding that both explicit and implicit memory are better when there is greater overlap between processes engaged during study and test (Morris et al., 1977; Blaxton, 1989). Thus, if the goal is to improve perception, perceptual training should be superior to production training. Another reason for why production may be detrimental to learning is errors. While errors can aid learning by triggering monitoring and control processes that help the system learn from error (i.e., error-based learning), errors themselves could be learned through production (Humphreys et al.., 2010; Waller et al., 2024). Finally, mixing production and perception can impose dual-tasking and attentional demands which could impede perceptual learning (Baese-Berk and Samuel, 2022; Kapnoula & Samuel, 2023). In short, there are good theoretical arguments on both sides for the possible effects of production on perception, many of which are not restricted by the nature of representations and stages of processing.

Empirical evidence for the influence of production on perception/comprehension is notoriously mixed. Some studies suggest that production during training helps recognition of newly trained vocabulary. For example, Zamuner et al. (2016) trained adults on eight new syllables paired with pictures. Four words were trained in the production mode, where participants heard and then repeated the word. The other four words were trained in comprehension mode, where participants simply heard the word twice. Analysis of eye-tracking data and a test probing recognition both showed better learning in the production mode. In another study, MacDonald (2018) trained young adults using an artificial language with sentences of increasing complexity. In the comprehension training mode, participants saw a scene, heard a phrase, indicated whether or not the phrase correctly described the scene, and received feedback. In the production training

mode, participants described the picture aloud, then heard the correct phase. Results showed faster and more accurate morphosyntactic knowledge in the production training mode. While there was no effect on accuracy, production training also led to faster responses times in a two-choice word-to-picture-matching task, taken as evidence for a positive influence of production on vocabulary learning.

However, these findings are not uncontested. For example, Zamuner et al. (2018) reported the opposite effect in 4.5-6-year-old children learning new vocabulary. In addition, a series of studies by Baese-Berk, Samuel, and colleagues have shown that learning acoustic differences between novel contrasts is more difficult when training includes both production and perception, as opposed to simple perception (Baese-Berk, 2010; 2019; Baese-Berk & Samuel, 2016; 2022), and that this difficulty extends to novel word recognition (Kapnoula & Samuel, 2022; 2023; Leach & Samuel, 2007). Interestingly, classroom-style studies of second-language acquisition also suggest that language production does not benefit language learning (Krashen, 2003; VanPatten, 2013; VanPatten & Cadierno, 1993).

Contextual facilitation and interference effects in language processing and learning

Contextual effects, i.e., facilitation or interference from other items in the set, are one of the most well-established findings in psycholinguistics. Two kinds of contextual relations are often studied: semantic relations investigate overlap in meaning, whereas segmental relations investigate overlap in form, phonemes in spoken production and graphemes in written production (see Nozari & Pinet, 2020, for a review). Interestingly, both kinds of relations could have facilitatory or inhibitory effects on language processing. For example, robust interference from the presence of a semantically related item has been obtained using picture-word interference paradigms, when a picture is named in the presence of a distracting word (e.g., Schriefers et al., 1990; see Bürki, et

al., 2020, for a meta-analysis), blocked cyclic naming paradigms, when a small set of pictures are repeatedly named (e.g., Belke et al., 2005; Damian et al., 2001; Schnur et al., 2006), and even simple continuous naming paradigms, when a series of pictures are simply named in a sequences (e.g., Costa et al., 2009; Hepner & Nozari, 2020; Schnur, 2014). Similar effects have been reported in language comprehension (e.g., Campanella & Shallice, 2011; Gardner et al., 2012; Biegler et al., 2008; Wei & Schnur, 2016).

At the same time, semantic facilitation effects have also been widely reported (e.g., Alario, et al., 2000; Costa et al., 2005; De Zubicaray et al., 2013; de Zubicaray et al., 2014; La Heij et al., 1990; Mahon et al., 2007). Some of the differences relate to timing. Short (< 2s) intervals between naming two semantically related pictures can facilitate naming (Biggs & Marmurek, 1990; Huttenlocher & Kubicek, 1983; Lupker, 1988), whereas longer (> 4 s) intervals seem to have the opposite effects (Vitkovitch et al., 2001; Wheeldon & Monsell, 1994). The type of semantic relation also appears to be important. Taxonomic relations (e.g., cat/dog) often induce interference, whereas non-taxonomic relations do not, and may even induce facilitation (Abel et al., 2009; Alario et al., 2000; Costa, Alario, & Caramazza, 2005; de Zubicaray et al., 2013; de Zubicaray et al., 2014; La Heij et al., 1990; Mahon et al., 2007; Oppenheim & Nozari, 2024; cf., Rose et al., 2019).

Similar contextual effects have been reported for segmental relations. Overlap in phonemes or graphemes between words (e.g., cup/map/cap) can cause interference in production (Breining et al., 2016; 2019; Feng et al., 2021; Harrison et al., 2020; Nozari et al., 2016; Pinet & Nozari, 2023; Qu et al., 2021; Sadat et al., 2014). However, segmental overlap can also be facilitatory. This is often the case when segmental relations make the next word predictable. For example, if the majority of items within a set start with a common onset (e.g., cap/cup/cat/…), participants can

strategically prepare that onset, which leads to its faster production (Nozari et al., 2016; O'Séaghdha & Frazer, 2014; Roelofs, 1999). Similar facilitation and interference effects have also been reported in comprehension (e.g., Radeau et al., 1989; Ziegler et al., 2003; Dufour, 2008).

Facilitatory effects of context are generally explained in terms of support received through shared features. It is now widely accepted that the activation of a target word in either comprehension or production entails the co-activation of other representations that share (semantic and/or segmental) features with the target word (Dell, 1984; Levelt et al., 1999). For example, when naming "dog", "cat" is also activated. The shared activation between dog and cat, makes "cat" easier to retrieve if it becomes the target on the next trial, a finding that forms the basis of positive priming. Since activation is generally short-lived, facilitatory effect of related primes are also usually fleeting (e.g., La Heij et al., 1990; cf., long-lasting repetition priming, e.g., Mitchell & Brown, 1988). In contrast, interference effects from related items can be surprisingly long-lasting. For example, semantic interference survives long temporal delays and many intervening items (Hepner & Nozari, 2020; Schnur, 2014). The longevity and resilience of interference effects have led to proposals that link them to implicit learning mechanisms (Howard et al., 2006; Oppenheim et al., 2010; Oppenheim & Nozari, 2024).

A prime example of a learning account of semantic interference is Oppenheim et al. (2010). In this model, upon naming a picture of "dog", "cat" is also activated through its shared semantic features with "dog". At the end of the trial, through error-based learning mechanisms, the system adjusts the connections between semantic features and the activated lexical items, with the goal of making the current target (i.e., "dog") easier to retrieve in the future. This means that connections between "dog" and its semantic features grow stronger, while at the same time, the connections between "cat" and the same semantic features grow weaker, thus making "cat" less accessible in

the system. Unlike activation-based priming, changes to connections are long-lasting, making interference effects the dominant effect over longer periods of time. Similar incremental learning mechanisms have been shown to underlie segmental interference (Breining et al., 2019; Qu et al., 2021).

The literature above is concerned with processing known words, but what about learning novel words? The incremental learning accounts, described above, suggest a direct link between online processing and long-term learning, by proposing that a continuous and incremental learning process underlies online processing. If so, one would expect the interference observed in online production and comprehension tasks to also affect the acquisition of new vocabulary (Oppenheim, 2018). Specifically, contextual similarity, semantic or phonological, between items in a training set should impede learning, compared to a set of unrelated items. This prediction is also compatible with some of the classic accounts of contextual effects in learning and memory, which posit that distinctiveness of a target facilitates its learning (the distinctiveness hypothesis; Hunt and Mitchell, 1982), and conversely, its increased similarity to other items interferes with its learning (the interference theory; Crowder, 1976).

In contrast to the theoretical accounts discussed above, which predict similar effects during online processing and long-term retention, some theories of learning and memory posit that greater difficulty in the processing of the materials during the study phase leads to better retention of information (e.g., Battig, 1972). However, not all difficulty during learning leads to better retention, causing researchers to use the term "desirable difficulties" to label those conditions which do improve long-term retention (Bjork 1994; Bjork & Bjork, 2011; Bjork & Kroll, 2015). The challenge is that identifying which difficulties are desirable and which are not, is not

straightforward (Bjork & Bjork, 2020). Therefore, it remains an empirical question, whether contextual similarity effects help or hurt the learning of new vocabulary.

If one consults the education literature, in addition to the cognitive literature, a non-trivial number of studies have examined contextual effects on learning. Some have reported facilitation in related contexts (e.g., Grandy, 2012; Hashemi & Gowdasiaei, 2005; Haycraft, 1978; Hoshino, 2010; Schnieder et al., 2002; Seal, 1991; Stoller & Grabe, 1993; Wharton & Race, 1999), and some interference (e.g. Breining et al., 2019; Finkbeiner & Nicol, 2003; Higa, 1963; Korochkina et al., 2021; Nation, 2000; Papagano & Vallar, 1992; Papathanasiou, 2009; Pérez-Serrano et al., 2022; Schneider et al., 1998, 2013; Tinkham, 1993, 1997; Waring, 1997; Wilcox & Medina, 2013). These studies have used very different populations and methodologies, including ways of measuring learning, methods of training (e.g., picture-word association vs. word-to-word translation), and level of control over the training materials (e.g., confounds of semantic similarity with phonological similarity), which makes comparison across them difficult. Two well-controlled cognitive studies in this literature have both reported contextual interference during learning. The first is Breining et al. (2019) in written production. The authors trained participants for four sessions on new labels and features of unfamiliar objects in semantically related, segmentally related and unrelated blocks. They found that both semantic and segmental similarity interfered with word learning. In a more recent study, Korochkina et al. (2021) trained young adults on novel labels for familiar objects in groups of semantically similar or unrelated items, in spoken production. Training took place over two sessions, 24 hours apart. Learning in the semantic condition was poorer throughout training, and immediately after the second session of training, the authors also found a detrimental effect of semantic similarity on response latencies in naming and translation tests, as well as greater semantic interference in a picture-word interference

paradigm. However, neither of these two studies probed the *comprehension* of new words. Also, neither study investigated long-term retention.

In summary, there are good theoretical arguments for both positive and negative effects of production training mode and contextual similarity on the comprehension of new vocabulary, but the evidence is mixed and there are missing pieces.

Current study

The aim of the current study was to investigate the effect of training mode and two types of contextual similarity on the comprehension of newly acquired vocabulary. In two experiments, we trained a total of 179 participants on 27 new vocabulary items for unfamiliar objects using an artificial language. Training mode and contextual similarity were each manipulated with three levels. Each participant was assigned either to a study (passive listening), a comprehension (listening and selecting), or a production (speaking) training mode. All participants completed blocks of semantically related, phonologically related, or unrelated items. We measured the learning of new vocabulary in comprehension using a word-to-picture-matching task, administered at different time points during the study.

Several changes and improvements were implemented over past studies: First, as in Breining et al. (2019), the current study included both semantic and segmental similarity within the same participants, allowing us to test contextual similarity effects in two stages of production in the same task and population. However, while Breining et al.'s study was in writing and probed learning in production, the current study was in speech and measured the comprehension of new vocabulary. If the detrimental effect of contextual similarity on learning is robust against these differences, we would expect to replicate the detrimental effects of both types of similarity in the present study. Second, mode and context were fully crossed in the design, allowing us to study the

potential interaction between the two factors. Recall that the bulk of evidence for the negative influence of production on perception came from studies targeting post-lexical differences (e.g., Baese-Berk & Samuel, 2016; 2022), whereas positive effects were reported in studies targeting semantic-lexical/syntactic mapping (e.g., Zamuner et al., 2016; Hopman & MacDonald, 2018). If the effect of mode critically depends on stage of processing, one would predict an interaction between mode and similarity, such that production is detrimental in the phonologically related, but facilitatory in semantically related context. Conversely, if the effect of training mode does not depend on stage of processing, the large sample size of the present study should allow us to observe a significant main effect of mode in one direction or another across all levels of contextual similarity.

Third, to ensure that differences between comprehension and production training mode were not attributable to differences in levels of engagement, both comprehension and production training modes incorporated active selection and feedback (c.f., Zamuner et al., 2016). We also included a passive study mode to track possible differences between passive and active modes, regardless of whether the active mode entailed production or not. Moreover, to avoid contamination from differences in syntactic comprehension, vocabulary training and testing were both conducted outside of sentential context (cf., Hopman & MacDonald, 2018). Fourth, our critical comprehension test (word-to-picture-matching) was repeated at multiple time points to assess the timeline of the effects. This included a delayed test, 48-72 hours after training in Experiment 2, after learning had had a chance to consolidate after the last training session.

Fifth, Experiment 2 was designed with two purposes in mind, to provide a conceptual replication of Experiment 1, and to test the effect of difficulty at the time of training on the two factors. In Experiment 1, all nine items within a block were included in a training cycle. In

Experiment 2, we broke each block into cycles of three items, making it easier to learn these triplets. Recall that the desirable difficulty accounts, discussed earlier, maintain that if the task is too difficult, learners will not benefit. Thus, any negative impacts of mode or contextual similarity in Experiment 1 may simply be due to the fact that the training task was too difficult. If so, we would expect the simplification in Experiment 2 to eliminate, or possibly reverse, negative impacts of mode and similarity.

Experiment 1

Method

Data and analysis files for both experiments can be found at the following Open Science framework repository: https://osf.io/x6byw/?view_only=3691c05c668447769fcf31cc53ce653f.

Participants

Multiple main effects are on interest in the current study, but for sample size estimation, we targeted the interaction between similarity and training mode, as interactions require larger sample to detect significant effects. No prior study, to our knowledge, has examined this interaction, therefore, we do not have a prior effect size and must set one a priori. Since our interest is in uncovering effects that are clinically meaningful, we set the effect size to medium (Cohen's $d = 0.5$) and conducted a power analysis using PANGEA (Version 0.2; Westfall, 2016), specifying a model with an interaction between one within subjects variable (similarity) and one between subjects variable (mode), each with three levels, and two random variables for participant and item. With 30 participants per training mode (N = 90) and 9 items per similarity condition, this simulation yielded 80.8% power to detect the critical interaction, and 82.8% to detect main effects.

Anticipating possible attritions, 98 participants were recruited online through Prolific. Participants were all native English speakers from the United States or Canada. Eight participants

did not complete the experiment due to technical issues. The remaining 90 participants (ages 20-40 years, mean age 30.74 years; 66.7% men, 31.1% women, and 2.2% non-binary) were randomly assigned to one of the three learning modes. The study was approved by Carnegie Mellon University's Institutional Review Board.

Materials

Materials consisted of six sets of images and three sets of words. 54 images were selected from Google Images and Novel Object & Unusual Name (NOUN) Database (Horst & Hout, 2014). The images were all colored images of unfamiliar objects to avoid conflict with pre-existing labels for familiar objects. These images were divided into six sets of nine (Appendix A). Three of the sets each formed a coherent semantic category: birds, flowers, and fruits. The other three sets did not form a coherent semantic category. One set was used for the practice trials at the start of the experiment, and the other two were randomly used for either the unrelated or the phonological similarity conditions. For all six sets, visual differences in color, orientation, and shape were balanced as much as possible. To test whether visual similarity was adequately controlled for, for each of the five image sets used in the experimental trials, we compared every possible pair of images by computing a Histogram of Oriented Gradients (HOG) vector for each image using scikit-image (version 0.24.0; van der Walt et al., 2014). We resized both images to a height and width of 300 pixels. Each image was processed by reading the file and converting it to grayscale using OpenCV (version 4.10.0.84; Bradski, 2000). The distance between two images was calculated as the norm of the difference between their HOG vectors, where if A and B are the same image, the distance is zero. A larger value indicates greater dissimilarity between images. For each image set, we then averaged the pairwise dissimilarity scores to get an overall comparison between the groups, as well as for the triplets used in Study 2. The dissimilarity

scores were 28.95, 31.62, 28.59, for the three semantic groups and 29.31 and 27.8, for the two unrelated groups, indicating that, overall, the semantic condition did not have greater visual similarity than the unrelated condition.

We created 27 novel words and divided them into three sets, one with phonologically related labels and the other two with unrelated labels to be assigned randomly to unrelated and semantic similarity conditions. Across sets, we balanced the number of syllables and phonemes (Appendix B). To quantify phonological similarity, we used position-independent phonological overlap, defined as the total number of phonemes shared by two strings, regardless of position, divided by the total number of phonemes in the two strings (Goldrick et al., 2010). Within each set, we computed position-independent phonological overlap scores first between all combinations of pairs, then averaged across all pairs in the set. Phonological similarity scores were 0.088 and 0.085 for the two unrelated sets, and 0.383 for the related set. This magnitude of difference in phonological overlap between unrelated and related sets was comparable to studies that have found clear interference effects in language production (e.g., Breining et al., 2016). The audio recordings were generated using Descript (https://www.descript.com/), an artificial voice program.

Label-image mappings were pseudo-randomly generated for each participant at the start of the experiment. For the semantic block, One of the three semantic image sets (birds, fruits, or flowers) was randomly paired with one of the unrelated label sets. For the phonological block, one of the unrelated image sets was randomly paired with the phonological label set. The remaining two unrelated sets then created the unrelated block. Within each block, the images were randomly mapped to labels for each participant. This procedure is important for ensuring that any idiosyncrasies in pictures, audio stimuli, or the pairing of the two is not contaminating the effects of interest. For consistency, practice trials were the same for everyone.

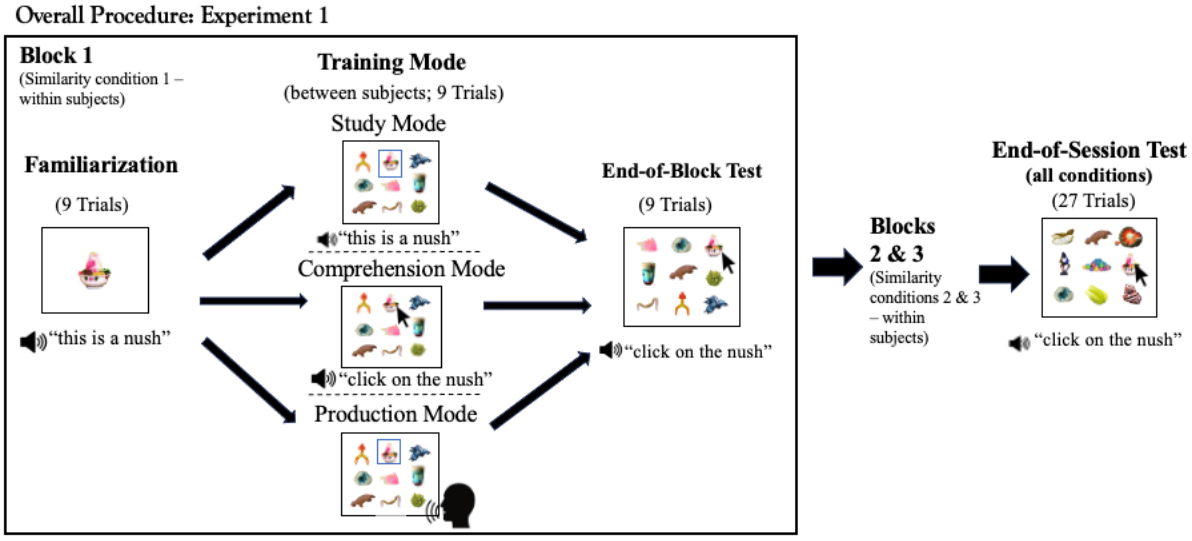

Figure 1: Overall procedure for Experiment 1. Similarity is manipulated within subjects across three blocks. Training Mode is manipulated between subjects and is consistent across blocks. The example is for an unrelated block. The other two blocks were semantically and phonologically related blocks.

Procedure

The task was conducted online using JavaScript code with JsPsych plugins. Participants were asked to learn novel labels for 27 pictures. Two sets of factors were manipulated, training mode and contextual similarity. Training mode consisted of three levels, *study, comprehension, and production*, and was manipulated between subjects. Contextual similarity also had three levels, *semantic, phonological, and unrelated*, but was manipulated within subjects. Participants were randomly assigned to one of the three training modes. Figure 1 shows the overall structure of a session. After general orientation and consenting, participants first completed three practice trials, the structure of which was similar to trials in the experimental blocks (see below). After practice, participants were presented with three blocks (semantic, phonological, and unrelated) in

randomized order. Each block consisted of a familiarization phase, a training phase, and an End-of-Block test. After completing all three blocks, all participants completed an End-of-Session test. The familiarization and test phases were identical across training modes, whereas the training phase was unique to each training mode.

In the familiarization phase, participants were introduced to each of the nine images in the block, one at a time, heard their labels (e.g., "This is a nush"), and were asked to repeat the word aloud before pressing a "continue" button to proceed to the next image. The purpose of including the repetition in this phase was to ensure that all participants paid attention to the familiarization phase and started the training phase on an equal footing. Note that repetition is not the production act deemed beneficial by proponents of production-mode training; rather it is the act of independently retrieving the item in the production system which has been argued to improve comprehension (Hopman & MacDonald, 2018). The order in which the images appeared was randomized for each participant. Next, participants moved on to training, in which each mode learned the labels differently depending on their training mode. As with familiarization, the order in which all nine items appeared was randomized for each participant. All trials began with a 3x3 grid of all nine images in a block, where the position of images changed on each trial. In the **study mode**, participants listened passively. Participants were first shown the grid, then after a 2000 ms delay a blue border appeared around the target image and the correct label was provided aloud (e.g. "this is a nush"). Then after another 2000 ms, the border turned green, and participants clicked on that image to proceed to the next trial. In the **comprehension mode**, 2000 ms after the 3x3 grid appeared, participants heard a label (e.g. "click on the nush") and had to select one of the nine images by clicking on it. When they selected an image, a blue border confirmed their selection, and then a green border appeared around the correct image to provide feedback. If the correct

image was the one they had selected, the blue border simply turned green. The task then automatically proceeded to the next trial. In the **production mode**, 2000 ms after the grid appeared, a blue border appeared around the target image, and participants were asked to say the name of that image. If they did not know the word, they were instructed to give their best guess. Once they decided they were done, they clicked on the image, and the audio of the correct label played aloud. After the audio played, the task automatically moved on to the next trial. Critically, we kept the 9-item grid constant for all training modes, in order to minimize visual differences between conditions. For all training modes, there was no response deadline, but after 5000 ms a reminder text appeared at the bottom of the screen to encourage them to respond and move on.

Immediately after completing training in each block, participants completed an **End-of-Block test**. This was a word-to-picture matching test to measure their learning in comprehension. Participants saw a 3x3 grid of all nine trained pictures, after 2000 ms heard one of the nine trained labels, and then selected the appropriate image using a mouse click. After the mouse click, a blue border appeared to confirm their selection, then after 1000 ms the task automatically proceeded to the next trial, where the grid was shuffled and a different label was heard. The order of trials was randomized for each participant. Similar to the training trials, there was no response deadline, but a reminder to respond appeared 5000 ms after the word was played. No feedback was provided in this phase. Finally, after completing all three blocks, participants completed the **End-of-Session Test**. The structure of this test was similar to the End-of-Block tests, except that participants completed 27 different trials, each with a 3x3 grid, this time with images from blocks mixed. The items that appeared in a grid on a given trial were pseudo-randomly determined, such that each of the 27 images appeared in the same number of trials (nine times), and there were always three from each block on the screen.

Data processing

Accuracy of comprehension selection and reaction times were automatically recorded using the JavaScript code. Audio files were transcribed by hand and coded for whether the participant accurately labeled the image, or whether they closely labelled the correct item (correctly using all but one phoneme), called Strict and Lenient accuracy respectively. The purpose of Lenient coding was to avoid penalizing participants for minor phonological/phonetic deviations when they had in fact retrieved the correct lexical item. Reaction times (RTs) were only analyzed for correct trials. RTs for any trials where the participant responded 3 standard deviations above or below their mean were further removed. RTs were log-transformed for the analysis.

Statistical analysis

Unless stated otherwise, all analyses were carried out using (general) linear mixed effect models with *lme4* package (Bates et al., 2015) in R (version 4.2.1, R Core Development Team, 2022). The goal of the analysis was to uncover potential effects of contextual similarity and training mode on vocabulary learning at different time points. For contextual similarity, the primary interest is to compare the effects of each type of similarity (semantic and phonological) against the baseline unrelated condition. Please note that a direct comparison of semantic and phonological conditions is not theoretically motivated, as the two dimensions of similarity are not comparable. For training mode, we were interested in the effect of the two active modes of training (comprehension and production) against the passive study mode, but we were also interested in directly comparing the two active modes against one another. To accommodate the three comparisons on mode, we ran two sets of models for each time point[1]. The first (main) model included contrasts of comprehension vs. study and production vs. study, whereas the second (direct comparison) model only included the subsets of data in the active training modes and compared

production directly against comprehension. All models coded contextual similarity as two contrasts, semantic vs. unrelated and phonological vs. unrelated. Following recommendations by Barr et al. (2013), we initially aimed to include the maximal random effect structure, and reduced the random effect structure if the models did not converge. For consistency we included the random intercept for subject and item, with which all models converged. The exact same structure was used for both accuracy and RT models, with the difference that a logistic version of the model was used for the latter with a binary accuracy measure. Results are organized chronologically, first reporting what is happening during training, mostly as a sanity check for the assumptions (e.g., production mode should be harder than comprehension mode), then End-of-Block test results, and finally End-of-Session test results. The full results of all models are reported in Appendix C.

Results

Figure 2 shows the accuracy during the within-block training phase across learning modes and contextual similarity, using lenient coding for production performance.[2] Note that participants in the study mode were passively listening, and therefore could make no errors. To check the starting assumption of the study, namely that production mode was more demanding than comprehension mode during training, we conducted a Mann-Whitney U test comparing participants' mean accuracy during training for these two modes. As expected, participants assigned to the production mode were less accurate than those assigned to the comprehension mode (M_{prod} = 23%, M_{comp} = 53%, W = 817, p < .001), validating our starting assumption.

Similarly, we checked the assumptions that semantic and phonological similarity conditions should impose greater difficulty during training compared to the unrelated condition. To this end, we conducted Wilcoxon Signed Ranks test comparing mean accuracy across participants for each similarity condition and the unrelated condition. This analysis only includes

participants from the comprehension and production modes, as only those participants made responses during training. For semantic similarity, there was a significant difference between performance, where participants performed worse in the semantic block compared to unrelated block (M_{Sem} = 28%, M_{Unrel} = 41% participants, V = 1029, p < .001). There was no significant difference between phonological and unrelated block accuracy (M_{Phon} = 44% M_{Unrel} = 41%, V = 409.5, p = .44).

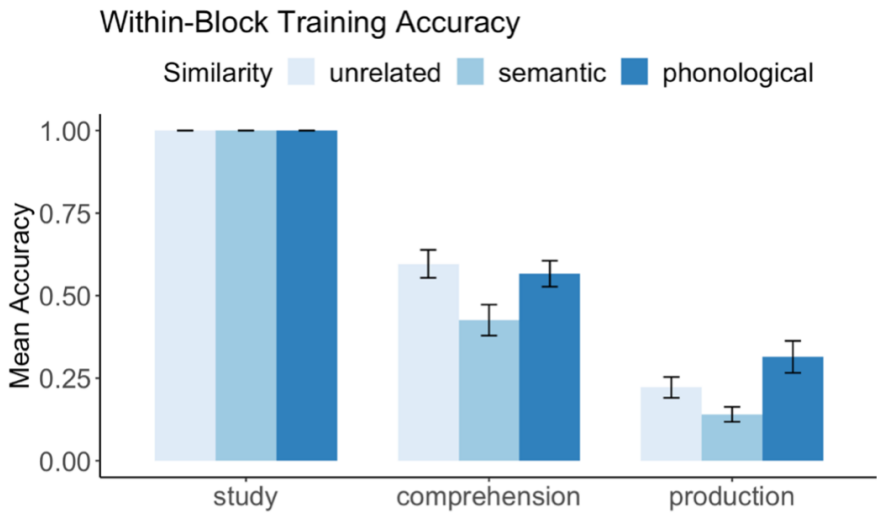

Figure 2. Average accuracy (with Lenient coding) during the within-block training trials in Experiment 1. Bars show mean of participant means, and error bars reflect SEs. Note that study participants could not make any errors during these trials. What is shown is clicks on the target picture, showing engagement.

End-of-Block tests

Figure 3 shows the accuracy (left) and RTs (right) on the End-of-Block tests for the three learning modes and contextual similarity. The accuracy model had mode, similarity and their interaction as its independent variables in the fixed effect structure and random intercept of

subjects and items as its random effects. Results are shown in Table C1. In terms of learning mode, accuracy in the production mode was marginally lower than study (z = -1.87, p = .06), but comprehension was comparable to study. In terms of similarity, while the main of effect of semantic similarity was not significant, both the interaction between comprehension mode and semantic similarity (z = - 3.49, p < .001) and production mode and semantic similarity (z = -2.481, p = .013) were significant, showing poorer accuracy in the semantic compared to unrelated similarity condition for both comprehension and production modes compared to study. Finally, neither the main effect of phonological similarity, nor its interaction with mode turned out to be statistically significant. To compare production and comprehension modes directly, we ran a second model with only data from these two modes included. Table C2 shows the results. Accuracy was significantly lower in the semantically related condition (z = -4.454, p < .001), but no other main effects or interactions were significant.

In the RT models, none of the main effects or interactions were significant in either model (Table C3 and C4), removing concerns regarding a potential speed-accuracy tradeoff.

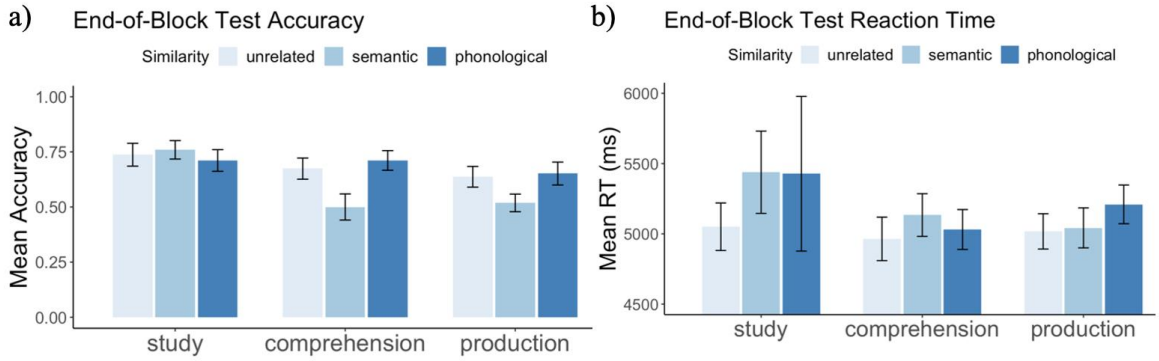

Figure 3. Average (a) Accuracy and (b) Reaction Time for End-of-Block comprehension tests in Experiment 1. Reaction time plots show mean reaction time on correct responses only. Bars show mean of participant means, and error bars reflect SEs.

End-of-Session test

Figure 4 shows the accuracy (left) and RTs (right) on the end-of-the-block tests for the three learning modes and contextual similarity. The model structures were identical to that used for analyzing the End-of-Block tests. The results of the accuracy analysis are shown in Table C5. There was a main effect of production (z = -2.436, p = .015), which did not interact with mode. None of the other effects on accuracy were significant. The model directly comparing production vs. comprehension mode again revealed that accuracy was significantly lower in production than comprehension (z = -2.356, p = .019), but also showed a main effect of semantic similarity with lower accuracy in semantically related compared to the unrelated condition (z = -2.498, p =.013; Table C6).

The results of RT analysis for the model comparing each of the comprehension and production modes to study are shown in Table C7. In this model, the interactions between comprehension (t = 2.072, p = .038) and production (t = 2.404, p = .016) modes and semantic similarity were both significant, suggesting slower responses in the semantic compared to unrelated similarity condition for both comprehension and production modes compared to study. As in previous models, neither the main effect of phonological similarity nor its interactions with mode were significant. The RT model directly comparing production and comprehension modes showed a main effect of semantic similarity, with longer RTs in the semantic compared to the unrelated condition (t = 3.57, p < .001). This model also revealed a significant interaction between production and phonological similarity, such that RTs were longer for phonological than unrelated condition in production compared to comprehension mode (t = 2.206, p = .028; Table C8).

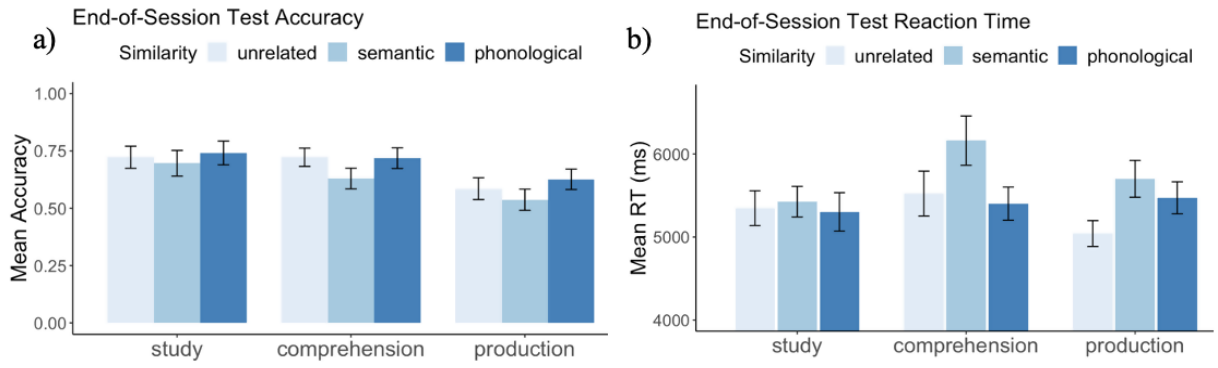

Figure 4. Average (a) Accuracy and (b) Reaction Time for End-of-Session comprehension test in Experiment 1. Reaction time plots show mean reaction time on correct responses only. Bars show mean of participant means, error bars reflect SEs.

Discussion

As expected, accuracy during training was lower in the production compared to the comprehension mode. Similarly, participants were less accurate while learning the labels in the semantically related, but not the phonologically related, condition. The End-of-Block test revealed less accurate performance on semantically related, compared to unrelated, blocks when embedded in production and comprehension, compared to study mode. When production was compared directly against comprehension, the model showed significantly lower accuracy for in semantically similar contexts. As for mode, accuracy in the production mode was marginally lower than study, but not significantly different from the comprehension mode.

The most critical test of the experiment, however, is the End-of-Session test, which measures learning of all trained items when probed in a mixed context. Here, there was clear evidence that learning in the production mode was less accurate than both the study mode and the comprehension mode. Semantic similarity also had a detrimental effect in these two active modes of learning: overall accuracy was lower in the semantic compared to the unrelated condition in the

model that included data from these two modes. Also, RTs were significantly slower in the semantic vs. unrelated condition for both production and comprehension modes vs. the study mode. The effect of phonological similarity on learning was much less robust. We only observed a disadvantage in RTs for learning in production compared to comprehension mode. To summarize, these results suggest a negative impact of the production mode on learning vocabulary in perception. They further show that semantic similarity among the items in the training set can be detrimental to learning in active learning modes.

Experiment 2 followed two goals. First, it was designed to provide a conceptual replication of Experiment 1; we tested whether the disadvantages observed for the production mode and semantic similarity were not due to general difficulty at the time of training. In Experiment 1, all nine items within a block were presented simultaneously to the participants. This choice led to relatively low performance during training, especially in the production mode (23%). However, many modern learning apps, e.g., Babble, train items in smaller clusters of 3 or 4, before moving on to the next cluster within the same session. Experiment 2 implemented this method within the general structure used in Experiment 1. A new group of participants were assigned to the same three modes as Experiment 1 and each completed training in three similarity contexts. However, within each block, items were trained as three triplets. We expected this change to lead to better within-block performance. The question was whether this change modifies the negative impact of production and semantic similarity on learning, or whether the observed effects are robust against such changes in the training routine. This manipulation is also important from a theoretical standpoint: if the negative impact of the examined factors is due to learning one's own errors, then reducing those errors in the training phase should, in turn, reduce the negative effects of production and similarity. The second goal of Experiment 2 was to test the longer-term effects of mode and

contextual similarity on learning, by being tested again 2-3 days after completing the training. This delayed assessment provides a further test of the robustness of the reported effects on learning.

Experiment 2

Method

Participants

Sample size estimation was the same as Experiment 1. Anticipating possible attrition, 101 participants were recruited online through Prolific. Participants were all native English speakers from the United States or Canada. Twelve participants did not complete the study, so their data were not included. The remaining 89 participants (ages 18-40 years, mean age 31 years, 50.6% men, 47.2% women, 1.1% nonbinary) were again randomly assigned to one of the three learning modes. Thirty participants completed the study and comprehension modes, and 29 completed the Production mode.

Materials

The same word and image sets as Experiment 1 were used, except one semantic image set (birds) was removed to accommodate changes in the word-image mapping procedure. Unlike in Experiment 1 where mappings were randomly determined for each participant, to ensure the mappings remained the same on both days of the experiment, participants were randomly assigned to one of four mapping lists, counterbalanced across modes. These lists balanced the matching of each word set with each image set, with the pairings of words and images randomly determined. For each of the four mappings, the label-image pairs within a block were created randomly. Participants were randomly assigned to one of these four mappings at the start of the experiment, counterbalanced across learning modes.

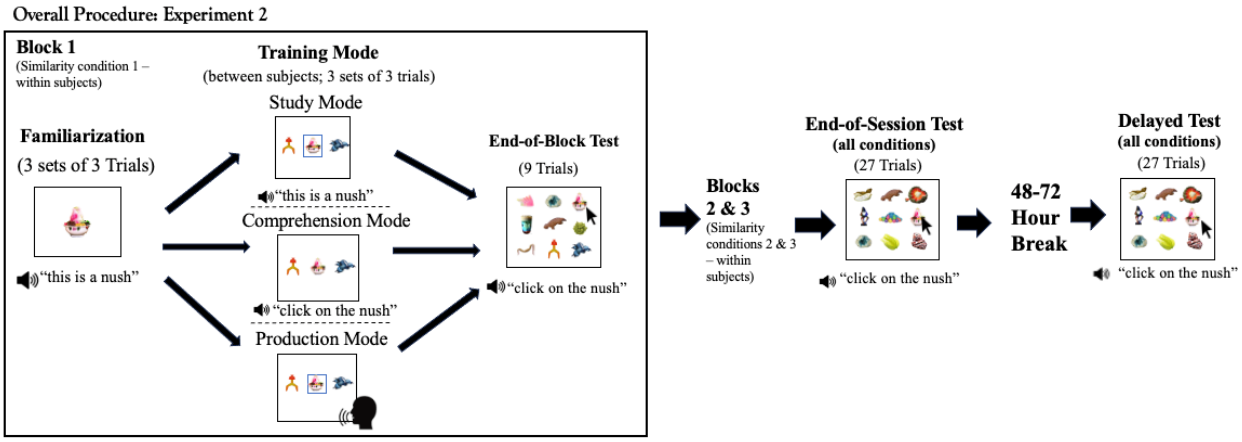

Figure 5. Overall procedure for Experiment 2. Note that Familiarization and Training occur in three sets of triplets, rather than with all 9 items at once. Participants complete Familiarization and training for one triplet before proceeding to the next triplet. All three triplets are completed before the End-of-Block test.

Procedure

As in Experiment 1, the task was conducted online using JavaScript code with JsPsych plugins. Figure 5 shows the procedure for this experiment. The general structure was similar to Experiment 1, with participants learning 27 novel words, in three modes (between subjects) and three similarity conditions (within subjects). However, Experiment 2 was different from Experiment 1 in two ways. The first difference from Experiment 1 was that within each block, rather than complete familiarization and training with all nine words at once, the nine items were divided into three triplets, and participants completed familiarization and training with each triplet separately. Which three images appeared together were pre-determined to control for visual similarity across sets. The order of the 3 triplets was randomized for each participant. After familiarization and training were completed on the triplets, participants completed an End-of-Block test identical to Experiment 1 containing all 9 items. After all three blocks were completed,

participants then completed an End-of-Session comprehension test that mixed together items from all three blocks, identical to Experiment 1. The second difference between Experiment 2 and Experiment 1 was the addition of a delayed test 48-72 hours after completing the first session. Participants completed another comprehension test identical to the End-of-Session test on the first session, with the 27 trials presented in randomized order.

Data processing and statistical analysis

Data processing and statistical analysis procedures were identical to Experiment 1.

Results

Figure 6 shows the accuracy during the within-block training phase across learning modes and contextual similarity, using Lenient coding for production performance.[3] We followed the same procedures for the verification of starting assumptions as Experiment 1, and the results were similar: there was a significant difference between modes, where participants were less accurate during the production mode compared to the comprehension mode (M_{prod} = 57%, M_{comp} = 89%, W = 813.5, $p < .001$). For semantic similarity, there was again a significant difference between performance, where participants performed worse in the semantic block compared to unrelated block (M_{Sem} = 69%, M_{Unrel} = 75% participants, V = 577.5, p = .023). There was no significant difference between phonological and unrelated block accuracy (M_{Phon} = 77% M_{Unrel} = 75%, V = 241, p = .226).

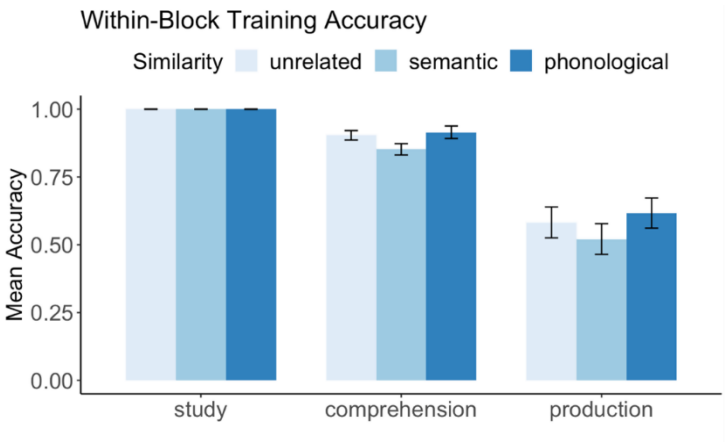

Figure 6. Accuracy (Lenient) during the within-block training trials in Experiment 2. Bars show mean of participant means, and error bars show SEs. Note that Study participants did not make any selection and therefore did not make any errors during these trials.

End-of-Block tests

Figure 7 shows the accuracy (left) and RTs (right) on the End-of-Block tests for the three learning modes and contextual similarity. The model structure was identical to the models used for analyzing Experiment 1. Results of the accuracy analysis are shown in Table C9. There was a significant main effect of semantic similarity (z = -3.217, p = .001), showing poorer accuracy for identifying items from the semantic similarity block across all modes. In the model comparing production and comprehension modes (Table C10) directly, there was also a significant main effect of semantic similarity (z = -4.135, p < .001). This model also showed a marginally poorer accuracy in production compared to comprehension (z = -1.919, p = .055). None of the other effects on accuracy were significant. The corresponding RT models (Tables C11 and C12) did not show any significant effects.

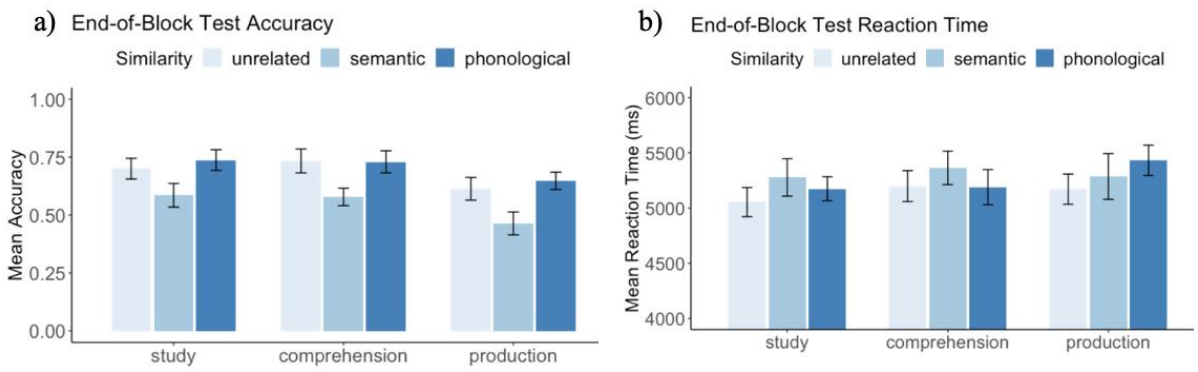

Figure 7. Average Accuracy (a) and Reaction Time (b) for End-of-Block comprehension tests in Experiment 2. Reaction time plots show mean reaction time on correct responses only. Bars show mean of participant means, error bars show SEs.

End-of-Session test

Figure 8 shows the accuracy (left) and RTs (right) on the Day 1 End-of-Session test for the three learning modes and contextual similarity. The model structures were identical to that used for analyzing the End-of-Block tests. Results of the accuracy analysis are shown in Table C13. There was again a significant main effect of semantic similarity (z = -2.35, p = .019), where participants performed worse on items from the semantic block across all three modes. There was also a significant main effect of production mode (z = -2.679, p = .007), where participants in the who completed production mode performed significantly worse across all similarity conditions compared to participants who were trained in the passive study mode. There were no other significant effects. In the model comparing production and comprehension modes (Table C14) directly, there was also a significant main effect of semantic similarity (z = -3.329, p = .001). This model also showed a marginally poorer accuracy in production compared to comprehension (z = -1.982, p = .047). The RT models (Tables C15 and C16) corresponding to both of the accuracy

models showed a robust disadvantage for semantic similarity (t = 3.756, p = < 0.001; t = 3.88, p < .001, respectively for the model comparing production and comprehension to study and for the direct model comparing production vs. comprehension). No other effects were significant.

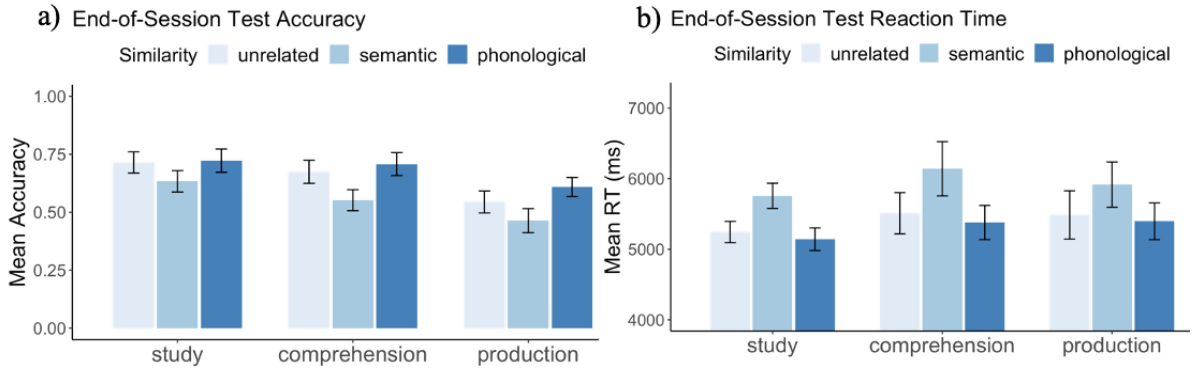

Figure 8. Average Accuracy (a) and Reaction Time (b) for End-of-Session comprehension test in on Day 1 in Experiment 2. Reaction time plots show mean reaction time on correct responses only. Bars show mean of participant means, error bars show SEs.

Delayed Test

Figure 9 shows the accuracy (left) and RTs (right) for the Delayed test conducted 48 to 72 hours after initial training. The results of the accuracy model are shown in Table C17. There were main effects of similarity for both semantic (z = -3.95, p < .001) and phonological (z = -2.221, p = .026) overlap. Participants across all modes showed poorer accuracy at identifying items trained in both similarity contexts compared to the unrelated context. Additionally, there was a main effect for the production mode (z = -2.068, p = .039), where participants who completed the production mode performed worse across all similarity conditions compared to participants who completed study mode. In the model comparing production and comprehension modes (Table C18) directly, there was again a significant main effect of semantic similarity (z = -4.921, p = < .001), as well as

a marginally poorer accuracy in production compared to comprehension (z = -1.982, p = .069). The RT models (Tables C19 and C20) corresponding to both of the accuracy models showed a robust disadvantage for semantic similarity (t = 4.281, p < .001; t = 4.253, p < .001, respectively for the model comparing production and comprehension to study and for the direct model comparing production vs. comprehension). Additionally, the RT models showed a marginally significant negative interaction between semantic similarity and production mode in the first model (t = -1.944, p = .052) and a significant negative interaction between the two in the model comparing production directly against comprehension (t = -2.177, p = .03), indicating a larger difference between production than comprehension modes for semantic compared to unrelated context.

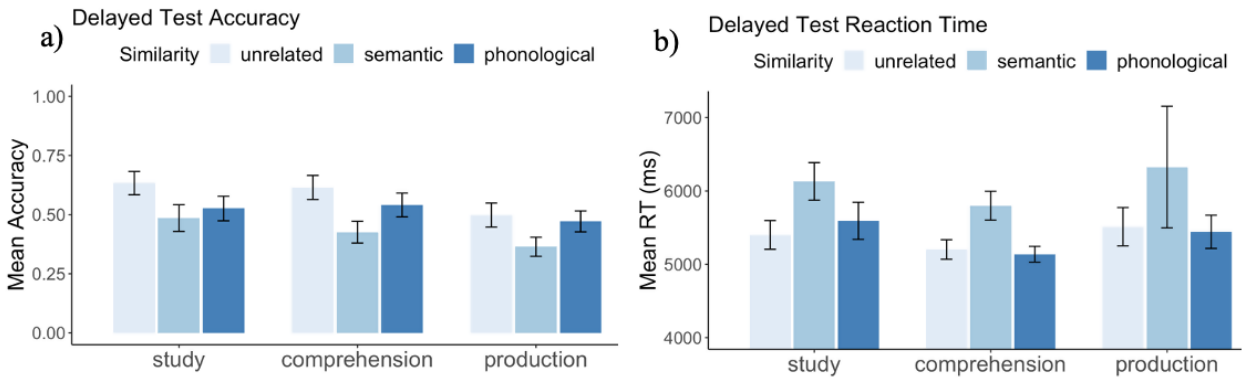

Figure 9. Average Accuracy (a) and Reaction Time (b) for Delayed comprehension test two days later in Experiment 2. Reaction time plots show mean reaction time on correct responses only. Bars show mean of participant means, error bars show SEs.

Discussion

Although, as expected, within-blocking learning was easier in Experiment 2, the results were largely similar to Experiment 1 and, in some ways, cleaner. For example, the detrimental effect of semantic similarity on the End-of-Block test in Experiment 1 was observed in the

interactions between production/comprehension and similarity in Experiment 1, but it showed up as a main effect in Experiment 2, suggesting a more general effect that held across all modes of learning. Importantly, the results of the End-of-Session testing were replicated: there was a robust detrimental effect of production compared to both study and comprehension modes. Moreover, while Experiment 1 only found a negative effect of semantic similarity on accuracy in the model that only contained production and comprehension data, there was again a main effect of accuracy across the board in Experiment 2, which was also reflected in slowed RTs in this similarity context compared to the unrelated context. Contrary to Experiment 1, however, there was no effect of phonological similarity on learning in the End-of-Session test.

Experiment 2 also tested the retention of information 48-72 hours after training. This test again supported the detrimental effect of semantic similarity on learning across the board. Moreover, learning was significantly worse in the production mode compared to study and marginally worse compared to comprehension. The Delayed test also showed a detrimental effect of phonological similarity, but only in the main model, and not the model that only included the production and comprehension data sets.

Combined Analysis of Experiments 1 and 2

Given the similarities between the designs of Experiments 1 and 2, we further conducted two sets of analyses on the combined data. The first analysis aimed at testing the effects of mode and contextual similarity with increased statistical power. With 179 participants, we have a power of 91.8% to detect main effects and of 85.9% to detect interaction effects with an effect size of Cohen's $d = 0.5$. This power also allows us to estimate more reliable size effects and to compare the effect size of production mode to study and comprehension directly. The second analysis was

planned to directly assess the effect of change of the influence of difficulty on mode and contextual similarity across the two experiments.

Analysis 1

Since Experiment 1 did not have a delayed test, this analysis was conducted on the End-of-Session test. The accuracy model had mode, similarity and the interaction between the two as the independent variables of interest. Experiment was entered as a covariate. We also included the random intercepts of subjects and items.[4] Table C21 shows the results for this model. There was no main effect of Experiment, showing overall comparable levels of accuracy across the two experiments, further supporting the pooling of the data. In the combined model, there was a main effect of semantic similarity (z = -2.277, p = .023), with poorer accuracy across all modes and across both experiments. Additionally, there was a main effect of production mode (z = -3.626, p < .001), as production participants performed worse across all contextual similarity conditions and across both experiments. The model directly comparing production and comprehension mirrored these results (Table C22): there was a significant negative impact of both semantic similarity (z = -4.092, p < .001) and production mode (z = -3.063, p = .002). To directly compare the effect sizes of the three levels of mode against each other, we used the emmeans package. Comprehension did not differ significantly from study (β = 0.31, z = 1.24, p = .429), but production showed significantly worse performance than study (β = 0.92, z = 3.76, p < .001) and comprehension (β = 0.62, z = 2.54, p = .030), with a larger effect size for the former. The corresponding RT models also found a significant negative effect of semantic similarity both in the first model (t = 3.285, p = .001, Table C23) and in the model comparing production directly to comprehension (t = 5.232, p < .001, Table C24). Other effects, including interactions, were not significant, suggesting largely

independent effects of semantic similarity and production mode on the acquisition of new vocabulary.

Analysis 2

Next, we ran the analyses focused specifically on testing whether difficulty during training (greater for Experiment 1 vs. Experiment 2) modulated the effect of mode or contextual similarity. A model with the three-way interaction of experiment, mode, and contextual similarity did not converge. Given that the previous analysis found largely independent effects of these two factors, we ran two separate analyses, one with experiment, mode and their two-way interaction, and another with experiment, contextual similarity, and their two-way interaction. General model structure and random effect structure were the same as the models reported earlier. For mode, we found significantly poorer accuracy in production compared to study mode (z = -2.766 p =.006; Table C25). The model comparing production directly to comprehension also revealed significantly poorer accuracy in production compared to comprehension mode (z = -2.095 p = 0.036; Table C26). There was no significant interaction in either model. Corresponding RT models did not show any significant effects (C27 and C28). For Contextual similarity, we found significantly poorer accuracy in the semantic compared to the unrelated condition (z = -2.663, p=.008; Table C29). The model comparing production directly to comprehension also revealed significantly poorer accuracy in the semantic compared to unrelated condition (z= -2.552; p= .011; Table C30). Again, there was no significant interaction in either model. Corresponding RT models also revealed significantly longer RTs for the semantic condition in both the main model (t = 4.414; p < .001; Table C31) and in the model comparing production vs. comprehension (t = 4.76; p < .001; Table 32). There were no other significant effects, including interactions. In short, there was

no evidence that difficulty during training modulated the effects of production mode and semantic similarity on the comprehension of new vocabulary.

General Discussion

In two experiments, we assessed the effects of contextual similarity and training mode on the comprehension of the new words in 179 adults. Despite differences in difficulty during the initial learning phase, the results of the two experiments were largely consistent, albeit with some minor differences. Significant negative effects of semantic similarity and marginal negative effects of production mode were apparent in early End-of-Block tests in both experiments. But the most important timepoints are the End-of-Session and Delayed tests, which measure immediate learning and long-term retention, respectively. For immediate learning, there was strong and convergent evidence from both experiments on detrimental effects of both production mode and semantic similarity on learning. When combined together, the data from the two experiments showed a clear negative impact of production mode compared to both study and comprehension modes, albeit with a larger effect size for the comparison to the less active study mode. Similarly, a significant negative effect of semantic similarity was evident across all three training modes. Importantly, with 179 participants in the combined dataset, we have a high statistical power to detect the medium-sized interaction between these two factors, but none of the interactions were significant, pointing to robust but largely separable effects of these two factors on vocabulary learning in comprehension. We also investigated the effect of difficulty during training on mode and similarity by testing the interactions between these variables and the two experiments. Again, we found no evidence that immediate learning was modulated by difficulty during training. Unlike semantic similarity, the effect of phonological similarity was sparse and weak and was not evident in the combined analysis. Therefore, our conclusion is that the immediate learning of new spoken

vocabulary in the comprehension system was negatively and independently impacted by production mode and semantic similarity, but not by phonological similarity among items in the training set.

Long-term retention, measured by the delayed test, was also negatively impacted by production and semantic similarity. However, the comparison of effect sizes for these two effects across immediate and delayed tests shows that the negative effect of semantic similarity on learning increased, while that of production mode decreased, with time. While a strong and pervasive effect of semantic similarity was present on both speed and accuracy in all models measuring long-term retention, learning was only significantly worse in production mode compared to the study, but not the comprehension, mode. Interestingly, the delayed test also showed a significant impact of phonological similarity, which did not interact with training mode. In summary, the results suggest that passage of time made the detrimental effects of both semantic and phonological similarity more prominent, whereas the negative impact of production mode was attenuated.

Comparisons with past studies and theoretical implications

The effect of training mode

The detrimental effect of the production mode on learning aligns well with studies of perceptual learning (Baese-Berk, 2010, 2019; Baese-Berk & Samuel, 2016, 2022; Kapnoula & Samuel, 2022; Krashen, 2003; Leach & Samuel, 2007; VanPatten & Cadierno, 1993; VanPatten, 2013; Zamuner et al., 2018), but not with those of Zamuner et al. (2016) and Hopman and MacDonald (2018) who reported better performance on comprehension tests for people who had engaged the production system in learning. One explanation for this differential alignment could be the inclusion of the phonologically similar block in our study, which required closer attention to sounds, something that was a focus of those studies that have reported interference from

production training, but not those that have reported facilitation. There was some support for this. In Experiment 1, the difference in the End-of-Session's RTs between production and comprehension modes was larger for the phonological, but not for the semantic, similarity condition. However, this effect was not replicated in Experiment 2, and there was no evidence of an interaction between production mode and phonological similarity on the delayed test, where both factors showed a negative main effect. Thus, although the negative impact of production on learning phonological/phonetic distinctions may be greater than semantic distinctions for lexical items, the current data do not provide strong support for the restriction of production effects to one stage of language processing.

Why the discrepancy with findings of Zamuner et al. (2016) and Hopman and MacDonald (2018)? One difference between Zamuner et al.'s (2016) study and ours was that their comprehension mode was passive, whereas ours required active engagement and selection. We included a passive control mode, but interestingly, learning was, for the most part, comparable between that mode and comprehension. One possible explanation for this finding is that participants were not paying attention in either study or comprehension modes, and thus did not learn much in either. The data do not support this interpretation. In fact, learning and retention were quite good in both study and comprehension modes, when compared to chance in a 9-item test (11%). A more likely interpretation is that even though the study mode did not elicit active retrieval, the use of the 9-item grid and the need for participants to click the highlighted image have focused attention sufficiently strongly for learning purposes. Another possibility is that passive studying is beneficial early on in the learning process, leading to similar performance between the two conditions, whereas active learning becomes more helpful later, after a baseline is established (MacDonald & Frank, 2016; Markant & Gureckis, 2014). Either way, the results

show that when the training mode is designed to keep the learner engaged during training, production does not lead to superior learning.

In contrast to Zamuner et al. (2016), Hopman and MacDonald's comprehension mode was engaging and required a choice between two options. But all testing in that study was carried out in the context of sentences. This is important, because the only measures that showed improved *accuracy* were morphosyntactic constructs, such as suffixes and their dependencies within a sentence. This finding suggests that the benefit of production training is most evident when grammatical parsing is needed. It is thus quite possible that the faster mapping of sentences to corresponding pictures, which was the metric used to assess the comprehension of new vocabulary, would owe much to participants' better developed abilities in parsing sentential grammar, as opposed to the meaning of individual lexical items. A direct comparison of the influence of training mode on isolated vs. within-sentence vocabulary learning is a great avenue for future research.

The absence of a positive effect of production training on vocabulary learning in comprehension seems to imply that production training is not a desirable difficulty. However, the decrease in the size of the detrimental effect of production mode on learning as a function of time may suggest otherwise. While memory for vocabulary naturally deteriorated in all groups after the delay, it deteriorated *less* for the production training group compared to their starting point, causing a reduction in the disadvantage observed in the production group compared to others over time. This finding, albeit too weak to be called a production "advantage", is aligned with the spirit of desirable difficulty: what is hard to learn in the beginning, is less susceptible to forgetting in the long run (Bjork & Bjork, 2011; Suzuki et al., 2019).

But why might production hurt learning in comprehension? One possibility is that participants in the production group are allowed to make errors, which they may learn instead of

correct responses. While there is some support for this in past studies (e.g., Humphreys et al., 2010; Waller et al., 2024), two findings from past studies speak against this idea as the sole explanation for the production disadvantage. First, Baese-Berk (2019) found no correlation between the number of production errors and the magnitude of disruption of perceptual learning. Our current results add to this finding, by showing that doubling production accuracy during training did not alter the negative effect of production. Second, in a clever manipulation, Baese-Berk and Samuel (2016) showed that *any* engagement of the production system, even if irrelevant to the targets of learning (i.e., reading an unrelated letter off the screen), caused some disruption to perceptual learning, although not to the level of attempting to produce the target of learning. This finding implies that the negative effect of production on perceptual learning does not entirely stem from target-related production attempts (e.g., errors). Rather it points to a more general effect of engaging the production system when the focus should be on perception, creating a dual-tasking effect. In keeping with this, Baese-Berk and Samuel (2022) and Kapnoula and Samuel (2023) showed that delaying production can reduce and ultimately eliminate its negative effects. While our study was not designed to replicate the delay condition, the current findings can help further define the attention account. Critically, we observed no change to the detrimental effect of production when the task was otherwise made easier, implying that the problem with production is not simply added general difficulty. The absence of a general effect of difficulty, combined with a local effect of delay at the item level, suggests that the problem is of the bottleneck type (Pashler, 1984), compatible with the interpretations of Baese-Berk and Samuel (2022).

At a broader level, these findings fit well with the classic theory of transfer-appropriate processing from the memory literature: memory performance is better when there is greater overlap in the processes engaged in training and test, as long as engagement and depth of

processing can be equated between different training modes. Note that the target processes here are not small task details, but larger operations and the general systems involved. In keeping with this, even though task details differed between study and comprehension modes, performance in these two conditions were largely comparable because the underlying system (the comprehension system) was engaged in both modes and in the test phase. Production mode, on the other hand, while still sharing the ultimate learning goal, involved processes that overlap less with retrieval in the comprehension system, causing a disadvantage.

The effect of contextual similarity

The interference induced by semantic similarity is well-aligned with the findings of Breining et al. (2019) and Krochkina et al. (2021) and adds to them by showing that this negative impact is not confined to learning in the *production* system, but also appears when tested in *comprehension*. Additionally, the current results extend the scope of influence of semantic similarity on learning to at least 2-3 days after the acquisition of the new vocabulary. Theoretically, these finding fit well with the incremental learning accounts of semantic interference (Oppenheim et al., 2010; Oppenheim & Nozari, 2024), and the previous reports on the longevity of semantic interference across long gaps and intervening items (Hepner & Nozari, 2020). They are also a good fit to distinctiveness hypothesis and interference theory, both of which also posit that learning should suffer as a consequence of increased similarity between items in the training set. These results, however, do not support semantic similarity as a "desirable difficulty". Not only was learning consistently worse in the semantically related context, but it also deteriorated faster (compared to the unrelated condition) as time passed. In contrast, desirable difficulty has often been argued, and shown, to especially benefit long-term retention (Bjork & Bjork, 2011; Suzuki, et al., 2019). One explanation for this finding could be that cognitive load was simply too high in

the semantic condition for participants to learn. If so, making the task easier by breaking training sets into small triplets should have ameliorated the detrimental effects, but we found no support for this. We, thus, propose that the detrimental effect of semantic similarity on vocabulary learning results from core implicit incremental processes which map semantic knowledge to lexical items (Oppenheim et al., 2010; Oppenheim & Nozari, 2024).

In contrast to semantic similarity, the effect of phonological similarity was less robust, and only affected accuracy on the delayed test. This is in contrast to the study of Breining et al. (2019), which found a robust effect of segmental similarity on learning. We offer two explanations for this difference: first, while there is now sufficient evidence to support the presence of phonological interference in production (e.g., Breining et al., 2016; Nozari et al., 2016; Harrison et al., 2020; Qu et al., 2021), the effect is more elusive than semantic interference, because it is more sensitive to strategies such as noticing common onsets (O'Séaghdha & Frazer, 2014). Such strategies can lead to short-term facilitation (Nozari et al., 2016), which may explain the absence of a robust effect in immediate training and its late emergence in the delayed test, when priming effects no longer apply. Second, Breining et al.'s (2019) study used written forms, both to study and to respond. As such, segmental similarity was not only evident in phonemes, but also in letters, the combined effect of which may have led to a stronger effect compared to the current study, which included no written forms. Clearly more research is needed to test the robustness of phonological interference in learning.

Clinical implications, limitations, and future directions

Linking cognitive principles to clinical settings always requires an intermediate link, namely larger-scale more ecologically valid studies that better imitate learning in real-life environments. However, well-controlled cognitive experiments provide valuable guidelines as to

which factors should be the target of such studies. Many training methods and language training apps contain sections for teaching vocabulary in isolation using word-picture mapping, similar to the current study. The same is true for many speech rehabilitation programs for individuals with language disorders due to brain damage, making our results relevant to current pedagogical and clinical practices. In most of these apps/programs, training items are arranged in semantically related sets. Our findings and others suggest that a taxonomic relationship between items can be detrimental to learning and the effect can be long-lasting, calling for large-scale trials to compare the effect of this factor on learning, e.g., in language apps. One usual caveat in studies of taxonomic similarity is the confound with visual similarity. While this problem is not limited to current study (e.g., see also Korochkina et al., 2021), we took care to match visual similarity as much as possible between semantically related and unrelated categories. However, in real life taxonomically similar items, especially from natural categories such as mammals, are generally more visually similar to one another than members of other categories, because physical features form an important basis for categorizing them together in the first place. Moreover, visual similarity is often stronger for the prototypical members of a natural category (e.g., cow and horse) vs. atypical ones (bat and aardvark), and language lessons often start with prototypical category members. Thus, acknowledging the correlation between taxonomic and visual similarity, instead of attempting to do away with it experimentally, is a more fruitful approach to understanding the best arrangement of pedagogical materials. On the other hand, visual similarity is not equally strong for *thematic* relations (e.g., balloon, cake, present). Incidentally, taxonomic and thematic similarity also affect production differently (Oppenheim & Nozari, 2024), therefore, the current results cannot be readily extended to thematically related sets. In fact, there is even reason, based on some facilitatory effects of thematic relations discussed earlier in the paper, that arranging materials

based on a theme, rather than a category, may benefit learning. An empirical investigation of this issue is a great avenue for future research.

Our findings together with those of past studies (e.g., Breining et al., 2019) also suggest that large-scale trials should investigate the influence of phonological/orthographic overlap among training items on learning. One issue here is the strength of segmental overlap. Note that semantic and segmental similarity are fundamentally different. Therefore, comparing the strengths of the two effects is meaningless. However, one may object that the less consistent effect of phonological, compared to semantic, similarity observed in the present study, may have been due to a weak manipulation. The strength of our manipulation was informed by two factors. First, this difference in overlap between phonological and unrelated conditions was selected to be comparable to prior studies (Breining et al., 2016; 2019) which had found interference effects in immediate production and in learning in the orthographic domain. Second, phonological similarity among labels in English is generally not very high. We, therefore, did not aim to construct a set with unrealistically high segmental similarity. That said, it is possible that this difference was simply too small to lead to more robust effects. A parametric manipulation of phonological and orthographic similarity to understand its effects on learning is another good topic for future research. Finally, in the present study, we manipulated semantic and phonological similarity independently to identify their unique influences on learning, but the two might interact. Examining the possible consequences of such an interaction is a good topic for future studies.

The question of training mode gains new importance in light of clinical recommendations. It is reasonable to restrict investigation to comprehension, if the question is a purely theoretical one, but in practice, no language learner would aim to only learn to understand, but not speak, the new language they are learning. Does this make the current findings irrelevant? We believe not.

The relevance may be in using the appropriate mode in different stages of training. Past work has shown that the detrimental effect of production training on learning new phonetic distinctions were tampered when the population of learners already had some experience with the learning materials (Baese-Berk and Samuel, 2016, Experiment 2). Our results add to this finding by showing that lowering task difficulty does not make up for lack of experience with materials. Naïve learners are hurt by production training mode even when training makes production much easier. Based on current findings, we would predict that restricting naïve learners' training to perception may give them a faster and stronger foothold on which they can then build production representations. Similarly, the benefit of production training on sentence processing as opposed to isolated vocabulary (e.g., Hopman & MacDonald, 2018) also implies that such training may be more efficient in later stages of learning when learners have acquired some basic vocabulary. Larger-scale studies investigating the addition of production training at different time-points in the course of learning can test this hypothesis.

Although we believe that the current findings are relevant to pedagogical practices, we acknowledge that no training study, including the current one, is perfect. Some limitations include the restriction of testing to a maximum of 72 hours post-training and the use of a limited set of concrete items, both of which limit a broader generalization of the results. Every training study must make choices among many experimental parameters that could further affect the results. These include, but are not limited to, using known vs. novel objects, cross-model (e.g., spoken + written) vs. unimodal training, different populations (e.g., children, younger adults or older adults), duration of training, etc. We thus aimed to reduce interference from current labels and synergistic effects of different modalities (the strength of both of which may differ across individuals), to focus on young adults whose linguistic systems are well-formed and not yet affected by cognitive

aging, and to obtain enough data to allow us to conduct statistical analyses with good power, without losing participants in a large multi-session study.

Furthermore, the comparison between some of the current and past design choices show that the effects are likely to be robust against these differences. For example, the detrimental influence of production on vocabulary learning cannot be attributed to age differences (c.f., Zamuner et al., 2016 vs. 2018), as our adult participants showed a pattern similar to Zamuner et al.'s (2018) child participants. Similarly, the number of training sessions or the use of known vs. novel objects do not seem to alter the negative influence of semantic similarity, as our findings mirror those of Krochkina et al. (2021), which differed from us in both factors. On the other hand, cross-model training, or a greater emphasis on one modality vs. another, does seem to affect the influence of segmental similarity, as that effect was more robust in a study with written representations (Breining et al., 2019), calling for more attention to this factor.

Conclusions

We found robust evidence for a detrimental and persistent effect of taxonomic similarity, and a less robust effect of phonological similarity, on comprehending newly acquired vocabulary, both of which grew more evident with the passage of time. These findings fit well with theories of incremental learning, which view online processing and long-term retention to be subject to similar principles. In addition, we found a negative and largely independent influence of production mode. However, unlike contextual similarity, the negative effect of production mode shrank with time, suggesting that it may stem from processes different from those underlying the negative influence of contextual similarity. Importantly, neither effect was reduced by decreasing difficulty at the time of training. This finding rules out certain explanations, such as the overall level of difficulty or the number of errors during training, as the culprit for the detrimental effects of production and

mode. Collectively, these results shed further light on the critical factors that influence new vocabulary learning and open new avenues for larger-scale studies to link them to pedagogical practices.

Acknowledgments

We thank Nikhil Lakhani for his assistance with building the experiment JavaScript code, and Anthony Tomasic for his assistance in implementing the HOG visual similarity algorithm. This work was supported by the James S. McDonnell Foundation Scholar Award in Understanding Human Cognition #220020506 to DY and NSF- BCS- 2346989 and Spencer Foundation grant 202000221 to N.N.

Declaration of interest statement

The authors report there are no competing interests to declare.

References

Abel, S., Dressel, K., Bitzer, R., Kümmerer, D., Mader, I., Weiller, C., & Huber, W. (2009). The separation of processing stages in a lexical interference fMRI paradigm. *Neuroimage*, *44*(3), 1113-1124.

Alario, F.-X., Segui, J., & Ferrand, L. (2000). Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(3), 741-764.

Baese-Berk, M. M. (2010). *An examination of the relationship between speech perception and production* (Doctoral dissertation, Northwestern University).

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, *81*, 981-1005.

Baese-Berk, M. M., Kapnoula, E. C., & Samuel, A. G. (2024). The relationship of speech perception and speech production: It's complicated. *Psychonomic Bulletin & Review*, 1-17.

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language, 89*, 23–36.

Baese-Berk, M. M., & Samuel, A. G. (2022). Just give it time: Differential effects of disruption and delay on perceptual learning. *Attention, Perception, & Psychophysics*, *84*(3), 960-980.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker., S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw*, *67*(1), 1-48.

Battig, W. F. (1972). Interference During Learning as a Sources of Facilitation in Subsequent Retention and Transfer.

Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology Section A*, *58*(4), 667-692.

Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech perception and linguistic experience*, *171*.

Biegler, K. A., Crowther, J. E., & Martin, R. C. (2008). Consequences of an inhibition deficit for word production and comprehension: Evidence from the semantic blocking paradigm. *Cognitive neuropsychology*, *25*(4), 493-527.

Biggs, T. C., & Marmurek, H. H. (1990). Picture and word naming: Is facilitation due to processing overlap?. *The American Journal of Psychology*, 81-100.

Bixby, K. N. (2017). *Production effects on perception: How learning to produce sound changes auditory perception*. University of Rochester.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 396 – 401). Wiley.

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp.185-205). MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American journal of psychology*, *128*(2), 241-252.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal ofApplied research in Memory and Cognition*, *9*(4), 475.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 657.

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience*, *22*(7), 1504-1529.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l: Long-term retention of learning in perception and production. *Perception & psychophysics*, *61*, 977-985.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Breining, B., Nozari, N., & Rapp, B. (2016). Does segmental overlap help or hurt? Evidence from blocked cyclic naming in spoken and written production. *Psychonomic bulletin & review*, *23*, 500-506.

Breining, B., Nozari, N., & Rapp, B. (2019). Learning in complex, multi-component cognitive systems: Different learning challenges within the same system. *Journal Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1093–1106.

Bürki, A., Elbuy, S., Madec, S., & Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language*, *114*, 104125.

Campanella, F., & Shallice, T. (2011). Manipulability and object recognition: is manipulability a semantic feature?. *Experimental Brain Research*, *208*, 369-383.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test?. *Psychonomic bulletin & review*, *13*(5), 826-830.

Costa, A., Alario, F. X., & Caramazza, A. (2005). On the categorical nature of the semantic interference effect in the picture-word interference paradigm. *Psychonomic Bulletin & Review*, *12*(1), 125-131.

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences*, *106*(50), 21442-21446.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*, *104*(3), 268.

Crowder, R.G. (1976). *Principles of Learning and Memory.* Lawrence Erlbaum.

de Zubicaray, G. I., Hansen, S., & McMahon, K. L. (2013). Differential processing of thematic and categorical conceptual relations in spoken word production. *Journal of Experimental Psychology: General*, *142*(1), 131.

de Zubicaray, G., Johnson, K., Howard, D., & McMahon, K. (2014). A perfusion fMRI investigation of thematic and categorical context effects in the spoken production of object names. *Cortex*, *54*, 135-149.

Damian, M. F., Vigliocco, G., & Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, *81*(3), B77-B86.

Dell, G. S. (1984). Representation of serial order in speech: evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 222.

Denes, P. B., & Pinson, E. (1993). *The speech chain*. Macmillan.

Dufour, S. (2008). Phonological priming in auditory word recognition: When both controlled and automatic processes are responsible for the effects. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *62*(1), 33.

Feng, C., Damian, M. F., & Qu, Q. (2021). Parallel processing of semantics and phonology in spoken production: Evidence from blocked cyclic picture naming and EEG. *Journal of Cognitive Neuroscience*, *33*(4), 725-738.

Feng, C., Damian, M. F., & Qu, Q. (2022). A joint investigation of facilitation and interference effects of semantic and phonological similarity in a continuous naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(8), 1193.

Fillingham, J. K., Sage, K., & Lambon Ralph, M. A. (2006). The treatment of anomia using errorless learning. *Neuropsychological rehabilitation*, *16*(2), 129-154.

Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied psycholinguistics*, *24*(3), 369-383.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of phonetics*, *14*(1), 3-28.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, *13*(3), 361-377.

Gardner, H. E., Lambon Ralph, M. A., Dodds, N., Jones, T., Ehsan, S., & Jefferies, E. (2012). The differential contributions of pFC and temporo-parietal cortex to multimodal semantic control: exploring refractory effects in semantic aphasia. *Journal of Cognitive Neuroscience*, *24*(4), 778-793.

Guenther, F. H. (2016). *Neural control of speech*. Mit Press.

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language*, *62*(2), 113-134.

Grandy, R. E. (2012). Semantic fields, prototypes, and the lexicon. In *Frames, fields, and contrasts* (pp. 103-122). Routledge.

Hashemi, M. R., & Gowdasiaei, F. (2005). An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC journal*, *36*(3), 341-361.

Harrison, W., Hepner, C., & Nozari, N. (2020). Is Segmental Interference Position-dependent?. In *CogSci* (Vol. 2020, pp. 681-687)

Haycraft, J. (1978). *An Introduction to English Language Teaching.* Longman.

Hepner, C. R., & Nozari, N. (2020). The dual origin of lexical perseverations in aphasia: Residual activation and incremental learning. *Neuropsychologia*, *147*, 107603.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature reviews neuroscience*, *13*(2), 135-145.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393-402.

Higa, M. (1963). Interference effects of intralist word relationships in verbal learning. *Journal of verbal learning and verbal behavior*, *2*(2), 170-175.

Hopman, E. W. M., & MacDonald, M. C. (2018). Production Practice During Language Learning Improves Comprehension. *Psychological Science*, *29*(6), 961–971.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior research methods*, *48*, 1393-1409.

Hoshino, Y. (2010). The Categorical Facilitation Effects on L2 Vocabulary Learning in a Classroom Setting. *RELC Journal*, *41*(3), 301–312.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213-1216.

Houde, J. F., and Jordan, M. I. (2002). Sensorimotor adaptation of speech I: compensation and adaptation. *J. Speech Lang. Hear. Res.* 45, 295–310.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition*, *100*(3), 464-482.

Humphreys, K. R., Menzies, H., & Lake, J. K. (2010). Repeated speech errors: Evidence for learning. *Cognition*, *117*(2), 151-165.

Hunt, R. R., & Mitchell, D. B. (1982). Independent effects of semantic and nonsemantic distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 81.

Huttenlocher, J., & Kubicek, L. F. (1983). The source of relatedness effects on naming latency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 486.

Icht, M., & Mama, Y. (2015). The production effect in memory: A prominent mnemonic in children. *Journal of Child Language*, *42*(5), 1102-1124.

Jacoby L. L., Shimizu Y., Daniels K. A., Rhodes M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. Psychonomic Bulletin & Review, 12, 852–857.

Kapnoula, E. C., & Samuel, A. G. (2022). Reconciling the contradictory effects of production on word learning: Production may help at first, but it hurts later. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(3), 394.

Kapnoula, E. C., & Samuel, A. G. (2023). Wait long and prosper! Delaying production alleviates its detrimental effect on word learning. *Language, Cognition and Neuroscience*, *38*(5), 724–744. https://doi.org/10.1080/23273798.2022.2144917

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772-775.

Kaushanskaya, M., & Yoo, J. (2011). Rehearsal effects in adult word learning. *Language and Cognitive Processes*, *26*(1), 121-148.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance.

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*(1), 269-319.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(9), 1297-1317.

Korochkina, M., Bürki, A., & Nickels, L. (2021). Apples and oranges: How does learning context affect novel word learning? *Journal of Memory and Language*, *120*, 104246. https://doi.org/10.1016/j.jml.2021.104246

Krashen, S. (2003). *Explorations in Language Acquisition and Use.* Heinemann.

La Heij, W., Dirkx, J., & Kramer, P. (1990). Categorical interference and associative priming in picture naming. *British Journal of Psychology*, *81*(4), 511-525.

Lawrence, C. O., Guitard, D., & Cowan, N. (2024). Short-term retention of words as a function of encoding depth. Memory & Cognition, 1-19.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive psychology*, *55*(4), 306-353.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, *22*(1), 1-38.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

Lupker, S. J. (1988). Picture naming: An investigation of the nature of categorical priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 444.

MacDonald, K., & Frank, M. C. (2016). When does passive learning improve the effectiveness of active learning?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 38).

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*(4), 390-395.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671.

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, 143(1), 94.

Middleton, A., Fritz, S. L., & Lusardi, M. (2015). Walking speed: the functional vital sign. *Journal of aging and physical activity*, *23*(2), 314-322.

Mitchell, D. B., & Brown, A. S. (1988). Persistent repetition priming in picture naming and its dissociation from recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 213.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of verbal learning and verbal behavior*, *16*(5), 519-533.

Murphy, T.K., Nozari, N. & Holt, L.L. (2023) Transfer of statistical learning from passive speech perception to speech production. *Psychonomic Bulletin & Rev*iew, *31*, 1193-1205. https://doi.org/10.3758/s13423-023-02399-8

Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*(2), 287–311.

Nation, I. S. P. (2000). Learning vocabulary in lexical sets: dangers and guidelines. *TESOL J.* 9, 6–10.

Nozari, N. (2020). Neural basis of word production. In L. R. Gleitman, A. Papafragou, & J. C. Trueswell (Eds.), *The Oxford Handbook of the Mental Lexicon.*

Nozari, N., Freund, M., Breining, B., Rapp, B., & Gordon, B. (2016). Cognitive control during selection and repair in word production. *Language, Cognition and Neuroscience*, *31*(7), 886–903.

Nozari, N., & Pinet, S. (2020). A critical review of the behavioral, neuroimaging, and electrophysiological studies of co-activation of representations during word production. *Journal of Neurolinguistics*, *53*, 100875.

Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, *106*(1), 528-537.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010) The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227-252.

Oppenheim, G. M. (2018). The paca that roared: Immediate cumulative semantic interference among newly acquired words. *Cognition*, *177*, 21-29.

Oppenheim, G. M., & Nozari, N. (2024). Similarity induced interference or facilitation in language production reflects representation, not selection. *Cognition.*

O'Séaghdha, P. G., & Frazer, A. K. (2014). The exception does not rule: attention constrains form preparation in word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 797.

Papathanasiou, E. (2009). An investigation of two ways of presenting vocabulary. *ELT journal*, *63*(4), 313-322.

Papagno, C., & Vallar, G. (1992). Phonological short-term memory and the learning of novel words: The effect of phonological similarity and item length. *The Quarterly Journal of Experimental Psychology*, *44*(1), 47-67.

Pashler, H. (1984). Processing stages in overlapping tasks: evidence for a central bottleneck. *Journal of Experimental Psychology: Human perception and performance*, *10*(3), 358.

Pérez-Serrano, M., Nogueroles-López, M., & Duñabeitia, J. A. (2022). Effects of semantic clustering and repetition on incidental vocabulary learning. *Frontiers in Psychology*, *13*, 997951.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329-347.

Pinet, S. & Nozari, N. (2023). Different electrophysiological signatures of similarity-induced and Stroop-like interference in language production. *Journal of Cognitive Neuroscience*, *35*(8), 1329-1349.

Qu, Q., Feng, C., & Damian, M. F. (2021). Interference effects of phonological similarity in word production arise from competitive incremental learning. *Cognition*, *212*, 104738.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radeau, M., Morais, J., & Dewier, A. (1989). Phonological priming in spoken word recognition: Task effects. *Memory & Cognition*, *17*(5), 525-535.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and cognitive processes*, *14*(2), 173-200.

Rose, S. B., Aristei, S., Melinger, A., & Abdel Rahman, R. (2019). The closer they are, the more they interfere: Semantic similarity of word distractors increases competition in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 753.

Saxton, M. (1997). The contrast theory of negative input. *Journal of child language*, *24*(1), 139-161.

Saxton, M., Kulcsar, B., Marshall, G., & Rupra, M. (1998). Longer-term effects of corrective input: An experimental approach. *Journal of child language*, *25*(3), 701-721.

Sadat, J., Martin, C. D., Costa, A., & Alario, F. X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive psychology*, *68*, 33-58.

Schnur, T. T. (2014). The persistence of cumulative semantic interference during naming. *Journal of Memory and Language, 75,* 27-44.

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language*, *54*(2), 199-227.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. Journal of Memory and Language, 46, 419–440.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2013). Contextual interference effects in foreign language vocabulary acquisition and retention. In *Foreign language learning* (pp. 77-90). Psychology Press.

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of memory and language*, *29*(1), 86-102.

Schuchard, J., & Middleton, E. L. (2018). The roles of retrieval practice versus errorless learning in strengthening lexical access in aphasia. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1700-1717.

Seal, B. D. (1991) Vocabulary learning and teaching *In Celce-Murcia, M.(Ed.) Teaching English as a_Second or Foreign Language, 296-311*. Boston, MA: Heinle & Heinle.

Stoller, F., & Grabe, W. (1993). Implications for L2 vocabulary acquisition and instruction from L1 vocabulary research. *Second language reading and vocabulary learning*, 24-45.

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, *103*(3), 713-720.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, *21*(3), 371-380.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second language research*, *13*(2), 138-163.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., & the scikit-image contributors. (2014). scikit-image: Image processing in Python. *PeerJ, 2*, e453. https://doi.org/10.7717/peerj.453

VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in second language acquisition*, *15*(2), 225-243.

VanPatten, B. (2013). Input processing. In *The Routledge handbook of second language acquisition* (pp. 268-281). Routledge.

Vitkovitch, M., Rutter, C., & Read, A. (2001). Inhibitory effects during object name retrieval: The effect of interval between prime and target on picture naming responses. *British journal of psychology*, *92*(3), 483-506.

Waller, M., Yurovsky, D, & Nozari N. (2024). Of mouses and mans: a test of errorless versus error-based learning in children. *Cognitive Science*, *48*(11), e70006.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. Journal of the Acoustical Society of America, *113*(2), 1033–1043.

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System*, *25*(2), 261-274.

Wei, T., & Schnur, T. T. (2016). Long-term interference at the semantic level: Evidence from blocked-cyclic picture matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 149.

Westfall, J. (2016). PANGEA: Power ANalysis for GEneral Anova designs. (https://jakewestfall.shinyapps.io/pangea/)

Wharton, S., & Race, P. (1999). *500 tips for TESOL: Teaching English to speakers of other languages*. Psychology Press.

Wheeldon, L. R., & Monsell, S. (1994). Inhibition of spoken word production by priming a semantic competitor. *Journal of memory and language*, *33*(3), 332-356.

Wilcox, A., & Medina, A. (2013). Effects of semantic and phonological clustering on L2 vocabulary acquisition among novice learners. *System*, *41*(4), 1056-1069.

Zamuner, T. S., Morin-Lessard, E., Strahm, S., & Page, M. P. (2016). Spoken word recognition of novel words, either produced or only heard during learning. *Journal of Memory and Language*, *89*, 55-67.

Zamuner, T. S., Strahm, S., Morin-Lessard, E., & Page, M. P. (2018). Reverse production effect: Children recognize novel words better when they are heard rather than produced. *Developmental Science*, *21*(4), e12636.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & cognition*, *38*(8), 995-1008.

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779-793.

Appendix A

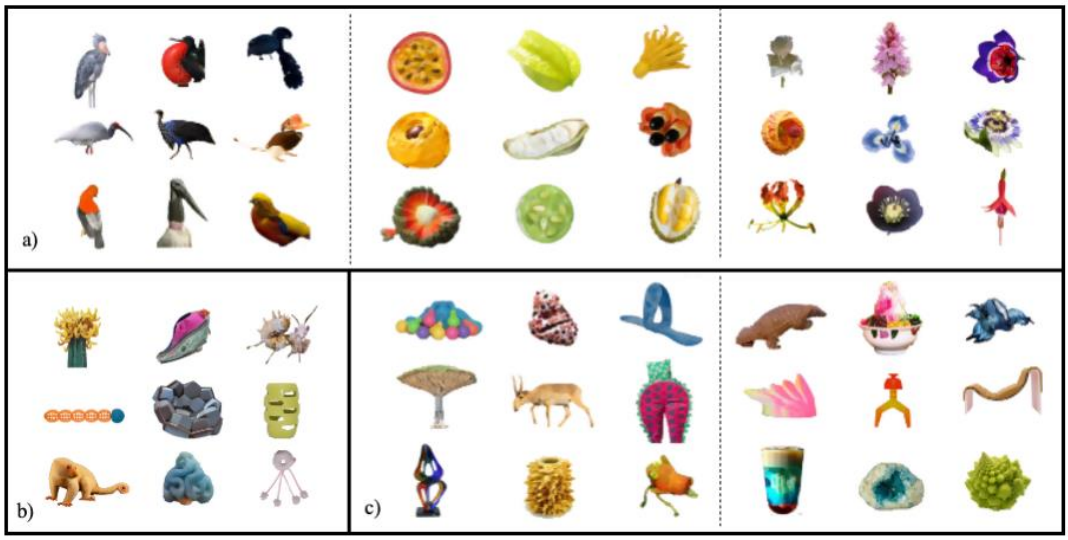

Figure A1: Novel image sets. a) The three sets of images with Semantic similarity. b) unrelated image set used for practice trials c) Two unrelated image sets used for training and test.

Appendix B

Unrelated Group 1

IPA	# of phonemes	# of syllables
lapi	4	2
hem	3	1
nʌʃ	3	1
dʒɪk	3	1
boʊlə	4	2
dɑrg	4	1
ræf	3	1
vɛŋ	3	1
mɑnɪt	5	2
average:	3.556	1.333

Unrelated Group 2

IPA	# of phonemes	# of syllables
sudl	4	2
goʊk	3	1
pæz	3	1
fʌp	3	1
teɪvoʊ	4	2
wɪlp	4	1
tʃim	3	1
zɛv	3	1
kibɛn	5	2
average:	3.556	1.333

Phonological Similarity Group

IPA	# of phonemes	# of syllables
rɑtu	4	2
rɑɪb	3	1
bɪs	3	1
sɑr	3	1
rɪsoʊ	4	2
bæst	4	1
sɛb	3	1
tɪb	3	1
tæsɛr	5	2
average:	3.556	1.333

Figure B1: IPA transcriptions used to calculate similarity averages, and their syllable and phoneme counts.

Appendix C

Exp 1

End-of-Block tests:

Table C1. Results of the accuracy analysis for the End-of-Block test in Exp 1, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.429	0.288	4.955	< 0.001
Comprehension	-0.463	0.386	-1.197	0.231
Production	-0.719	0.385	-1.871	0.061
Semantic	0.142	0.223	0.638	0.524
Phonological	-0.177	0.251	-0.706	0.48
Comprehension: Semantic	-1.051	0.301	-3.49	< 0.001
Production: Semantic	-0.738	0.298	-2.481	0.013
Comprehension: Phonological	0.402	0.3	1.338	0.181
Production: Phonological	0.265	0.295	0.898	0.369

Table C2. Results of the accuracy analysis for the End-of-Block test in Exp 1, comparing production vs. comprehension modes.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	0.947	0.26	3.645	< 0.001
Production	-0.247	0.35	-0.706	0.48
Semantic	-0.897	0.201	-4.454	< 0.001
Phonological	0.222	0.239	0.926	0.354
Production: Semantic	0.308	0.28	1.101	0.271
Production: Phonological	-0.135	0.286	-0.473	0.637

Table C3. Results of the RT analysis for the End-of-Block test in Exp 1, comparing production and comprehension modes against study

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.510	0.029	293.712	<0.001
Comprehension	-0.019	0.039	-0.489	0.626
Production	-0.016	0.039	-0.415	0.679
Semantic	0.037	0.023	1.618	0.106
Phonological	-0.002	0.029	-0.066	0.948
Comprehension: Semantic	-0.017	0.035	-0.494	0.622
Production: Semantic	-0.021	0.035	-0.605	0.546
Comprehension: Phonological	0.004	0.033	0.121	0.904
Production: Phonological	0.034	0.034	1.001	0.317

Table C4. Results of the RT analysis for the End-of-Block test in Exp 1, comparing production vs. comprehension modes

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.490	0.026	328.169	< 0.001
Production	0.004	0.034	0.122	0.903
Semantic	0.023	0.026	0.894	0.372
Phonological	0.003	0.029	0.098	0.923
Production: Semantic	-0.004	0.036	-0.107	0.915
Production: Phonological	0.030	0.033	0.890	0.374

Exp 2

Table C9. Results of the accuracy analysis for the End-of-Block test in Exp 2, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.111	0.261	4.254	< 0.001
Comprehension	0.145	0.352	0.41	0.681
Production	-0.492	0.348	-1.414	0.157
Semantic	-0.654	0.203	-3.217	0.001
Phonological	0.222	0.252	0.881	0.378
Comprehension: Semantic	-0.2	0.289	-0.692	0.489
Production: Semantic	-0.124	0.282	0.441	0.66
Comprehension: Phonological	-0.236	0.3	-0.786	0.432
Production: Phonological	-0.067	0.289	-0.233	0.816

Table C10. Results of the accuracy analysis for the End-of-Block test in Exp 2, comparing production vs. comprehension modes.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.241	0.249	4.985	< 0.001
Production	-0.633	0.33	-1.919	0.055
Semantic	-0.846	0.205	-4.135	< 0.001
Phonological	-0.018	0.249	-0.074	0.941
Production: Semantic	0.075	0.284	0.263	0.792
Production: Phonological	0.17	0.289	0.587	0.557

Table C11. Results of the RT analysis for the End-of-Block test in Exp 2, comparing production and comprehension modes against study

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.501	0.024	348.697	< 0.001
Comprehension	0.023	0.033	0.682	0.496
Production	0.016	0.034	0.463	0.644
Semantic	0.039	0.025	1.573	0.116
Phonological	0.022	0.026	0.855	0.394
Comprehension: Semantic	-0.002	0.035	-0.045	0.964
Production: Semantic	-0.003	0.038	-0.078	0.938
Comprehension: Phonological	-0.011	0.033	-0.328	0.743
Production: Phonological	0.033	0.035	0.928	0.354

Table C12. Results of the RT analysis for the End-of-Block test in Exp 2, comparing production vs. comprehension modes

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.525016	0.02534	336.424	< 0.001
Production	-0.008342	0.03484	-0.239	0.811
Semantic	0.036716	0.02505	1.466	0.143
Phonological	0.010953	0.02772	0.395	0.694
Production: Semantic	0.001665	0.03802	0.044	0.965
Production: Phonological	0.045055	0.03481	1.294	0.196

Exp 1 | Exp 2 | Combined test

End-of-Session tests:

Table C5. Results of the accuracy analysis for the End-of-Session test in Exp 1, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.387	0.294	4.721	< 0.001
Comprehension	-0.141	0.402	-0.351	0.726
Production	-0.965	0.396	-2.436	0.015
Semantic	-0.178	0.221	-0.808	0.419
Phonological	0.132	0.24	0.549	0.583
Comprehension: Semantic	-0.342	0.301	-1.136	0.256
Production: Semantic	-0.058	0.292	-0.198	0.843
Comprehension: Phonological	-0.149	0.307	-0.483	0.629
Production: Phonological	0.08	0.297	0.27	0.787

Table C6. Results of the accuracy analysis for the End-of-Session test in Exp 1, comparing production vs. comprehension modes

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.194	0.246	4.856	< 0.001
Production	-0.785	0.333	-2.356	0.019
Semantic	-0.506	0.203	-2.498	0.013
Phonological	-0.016	0.224	-0.072	0.943
Production: Semantic	0.277	0.277	1.001	0.317
Production: Phonological	0.225	0.283	0.796	0.426

Table C7. Results of the RT analysis for the End-of-Session test in Exp 1, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.533	0.033	254.923	< 0.001
Comprehension	0.035	0.044	0.791	0.430
Production	-0.029	0.045	-0.640	0.523
Semantic	0.018	0.025	0.703	0.482
Phonological	0.000	0.031	0.000	1.000
Comprehension: Semantic	0.073	0.035	2.072	0.038
Production: Semantic	0.090	0.037	2.404	0.016
Comprehension: Phonological	-0.011	0.034	-0.308	0.758
Production: Phonological	0.070	0.036	1.943	0.052

Table C8. Results of the RT analysis for the End-of-Session test in Exp 1, comparing production vs. comprehension.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.568	0.034	248.500	<0.001
Production	-0.064	0.046	-1.381	0.170
Semantic	0.091	0.025	3.570	<0.001
Phonological	-0.011	0.033	-0.351	0.727
Production: Semantic	0.016	0.037	0.422	0.673
Production: Phonological	0.079	0.036	2.206	0.028

Table C13. Results of the accuracy analysis for the End-of-Session test in Exp 2, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.263	0.288	4.383	< 0.001
Comprehension	-0.298	0.38	-0.785	0.432
Production	-1.011	0.378	-2.679	0.007
Semantic	-0.493	0.21	-2.35	0.019
Phonological	0.06	0.273	0.221	0.825
Comprehension: Semantic	-0.194	0.293	-0.661	0.508
Production: Semantic	0.082	0.286	0.285	0.776
Comprehension: Phonological	0.161	0.302	0.532	0.595
Production: Phonological	0.269	0.292	0.923	0.356

Table C14. Results of the accuracy analysis for the End-of-Session test in Exp 2, comparing production vs. comprehension modes

Fixed Effect	Coefficient	SE	z	p-value
Intercept	0.952	0.271	3.517	< 0.001
Production	-0.703	0.355	-1.982	0.047
Semantic	-0.679	0.204	-3.329	0.001
Phonological	0.221	0.269	0.819	0.413
Production: Semantic	0.266	0.283	0.941	0.347
Production: Phonological	0.106	0.288	0.367	0.367

Table C15. Results of the RT analysis for the End-of-Session test in Exp 2, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.528	0.035	245.654	< 0.001
Comprehension	0.026	0.048	0.529	0.598
Production	0.036	0.05	0.715	0.476
Semantic	0.102	0.027	3.756	< 0.001
Phonological	-0.018	0.029	-0.642	0.522
Comprehension: Semantic	0.014	0.039	0.363	0.716
Production: Semantic	-0.028	0.042	-0.672	0.502
Comprehension: Phonological	0.024	0.037	0.65	0.516
Production: Phonological	0.008	0.04	0.198	0.843

Table C16. Results of the RT analysis for the End-of-Session test in Exp 2, comparing production vs. comprehension.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.555	0.04	216.491	< 0.001
Production	0.009	0.055	0.155	0.877
Semantic	0.115	0.03	3.88	< 0.001
Phonological	0.006	0.033	0.168	0.868
Production: Semantic	-0.036	0.045	-0.793	0.428
Production: Phonological	-0.015	0.041	-0.354	0.724

Table C21. Results of the accuracy analysis for the End-of-Session test in both Exp 1 and Exp 2, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.45	0.237	6.117	< 0.001
Experiment	-0.246	0.2	-1.229	0.219
Comprehension	-0.223	0.276	-0.809	0.418
Production	-0.991	0.273	-3.626	< 0.001
Semantic	-0.346	0.152	-2.277	0.023
Phonological	0.1	0.209	0.481	0.63
Comprehension: Semantic	-0.255	0.21	-1.215	0.224
Production: Semantic	0.027	0.204	0.131	0.896
Comprehension: Phonological	0.009	0.216	0.043	0.966
Production: Phonological	0.176	0.208	0.847	0.397

Table C22. Results of the accuracy analysis for the End-of-Session test in both Exp 1 and Exp 2, comparing production vs. comprehension modes

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.193	0.222	5.37	< 0.001
Experiment	-0.233	0.214	-1.089	0.276
Production	-0.747	0.244	-3.063	0.002
Semantic	-0.588	0.144	-4.092	< 0.001
Phonological	0.11	0.208	0.528	0.597
Production: Semantic	0.271	0.198	1.37	0.171
Production: Phonological	0.164	0.202	0.81	0.418

Table C23. Results of the RT analysis for the End-of-Session test in both Exp 1 and Exp 2, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.524	0.028	305.913	< 0.001
Experiment	0.013	0.024	0.536	0.593
Comprehension	0.031	0.033	0.948	0.344
Production	0.002	0.033	0.061	0.952
Semantic	0.060	0.018	3.285	0.001
Phonological	-0.008	0.025	-0.311	0.757
Comprehension: Semantic	0.043	0.026	1.645	0.100
Production: Semantic	0.035	0.028	1.243	0.214
Comprehension: Phonological	0.006	0.025	0.256	0.798
Production: Phonological	0.042	0.027	1.569	0.117

Table C24. Results of the RT analysis for the End-of-Session test in both Exp 1 and Exp 2, comparing production vs. comprehension.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.557	0.032	269.528	< 0.001
Experiment	0.010	0.032	0.322	0.748
Production	-0.029	0.036	-0.820	0.414
Semantic	0.101	0.019	5.232	< 0.001
Phonological	-0.004	0.028	-0.128	0.899
Production: Semantic	-0.006	0.029	-0.212	0.832
Production: Phonological	0.036	0.027	1.323	0.186

Exp 1

Delayed test:

Exp 2

Table C17. Results of the accuracy analysis for the Delayed test in Exp 2, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	0.725	0.265	2.74	0.006
Comprehension	-0.104	0.348	-0.299	0.765
Production	-0.719	0.348	-2.068	0.039
Semantic	-0.789	0.2	-3.95	< 0.001
Phonological	-0.574	0.258	-2.221	0.026
Comprehension: Semantic	-0.194	0.282	-0.69	0.49
Production: Semantic	0.105	0.281	0.373	0.709
Comprehension: Phonological	0.188	0.28	0.669	0.503
Production: Phonological	0.436	0.278	1.568	0.117

Table C18. Results of the accuracy analysis for the Delayed test in Exp 2, comparing production vs. comprehension modes

Fixed Effect	Coefficient	SE	z	p-value
Intercept	0.616	0.256	2.405	0.016
Production	-0.607	0.334	-1.82	0.069
Semantic	-0.981	0.199	-4.921	< 0.001
Phonological	-0.385	0.262	-1.469	0.142
Production: Semantic	0.286	0.281	1.016	0.31
Production: Phonological	0.244	0.276	0.883	0.378

Table C19. Results of the RT analysis for the Delayed test in Exp 2, comparing production and comprehension modes against study.

Fixed Effect	Coefficient	SE	t-value	p-value
Intercept	8.562	0.034	255.497	< 0.001
Comprehension	-0.021	0.045	-0.477	0.634
Production	0.028	0.047	0.591	0.555
Semantic	0.12	0.028	4.281	< 0.001
Phonological	0.036	0.033	1.068	0.289
Comprehension: Semantic	-0.035	0.04	-0.857	0.391
Production: Semantic	-0.084	0.043	-1.944	0.052
Comprehension: Phonological	-0.048	0.038	-1.244	0.214
Production: Phonological	-0.05	0.041	-1.234	0.217

Table C20. Results of the RT analysis for the Delayed test in Exp 2, comparing production vs. comprehension.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.605	0.031	273.73	< 0.001
Production	0.035	0.043	0.808	0.421
Semantic	0.091	0.021	4.253	< 0.001
Phonological	0.015	0.026	0.553	0.582
Production: Semantic	-0.067	0.031	-2.177	0.03
Production: Phonological	-0.001	0.031	-0.035	0.972

Combined test

End-of-Session Test:

Table C25. Results of the accuracy analysis for the End-of-Session test in both Exp 1 and Exp 2, comparing production and comprehension modes against study and their interaction with experiment.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.373	0.260	5.272	< 0.001
Experiment	-0.261	0.348	-0.750	0.454
Comprehension	-0.306	0.348	-0.880	0.379
Production	-0.955	0.345	-2.766	0.006
Experiment : Comprehension	-0.014	0.488	-0.028	0.977
Experiment: Production	0.066	0.487	0.135	0.892

Table C26. Results of the accuracy analysis for the End-of-Session test in both Exp 1 and Exp 2, comparing production vs. comprehension modes and their interaction with experiment.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.039	0.228	4.561	< 0.001
Experiment	-0.264	0.301	-0.878	0.380
Production	-0.627	0.299	-2.095	0.036
Experiment: Production	0.067	0.424	0.158	0.875

Table C27. Results of the RT analysis for the combined test in both Exp 1 and Exp 2, comparing production and comprehension modes against study and their interaction with experiment.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.539	0.031	278.098	< 0.001
Experiment	0.016	0.042	0.390	0.697
Comprehension	0.054	0.042	1.297	0.196
Production	0.025	0.042	0.593	0.554
Experiment: Comprehension	-0.017	0.059	-0.294	0.769
Experiment : Production	0.005	0.060	0.091	0.927

Table C28. Results of the RT analysis for the End-of-Session test in both Exp 1 and Exp 2, comparing production vs. comprehension modes and their interaction with experiment.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.593	0.033	258.369	< 0.001
Experiment	-0.003	0.045	-0.062	0.951
Experiment : Comprehension	-0.030	0.045	-0.655	0.514
Experiment: Production	0.025	0.064	0.384	0.702

Table C29. Results of the accuracy analysis for all participants in the End-of-Session test in both Exp 1 and Exp 2, comparing semantic and phonological similarity conditions against the unrelated condition and their interaction with experiment.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	1.013	0.182	5.577	< 0.001
Experiment	-0.188	0.229	-0.821	0.411
Semantic	-0.314	0.118	-2.663	0.008
Phonological	0.127	0.184	0.692	0.489
Experiment : Semantic	-0.209	0.166	-1.256	0.209
Experiment : Phonological	0.085	0.170	0.499	0.618

Table C30. Results of the accuracy analysis for comprehension and production particiapants in the End-of-Session test in both Exp 1 and Exp 2, comparing semantic and phonological similarity conditions against the unrelated condition and their interaction with experiment.

Fixed Effect	Coefficient	SE	z	p-value
Intercept	0.804	0.195	4.122	< 0.001
Experiment	-0.211	0.249	-0.847	0.397
Semantic	-0.356	0.140	-2.552	0.011
Phonological	0.122	0.204	0.601	0.548
Experiment : Semantic	-0.180	0.198	-0.909	0.363
Experiment : Phonological	0.151	0.202	0.750	0.454

Table C31. Results of the RT analysis for all participants in the End-of-Session test in both Exp 1 and Exp 2, comparing semantic and phonological similarity conditions against the unrelated condition and their interaction with experiment.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.537	0.022	396.003	< 0.001
Experiment	0.011	0.027	0.402	0.688
Semantic	0.069	0.016	4.414	< 0.001
Phonological	0.017	0.023	0.760	0.451
Experiment : Semantic	0.033	0.023	1.473	0.141
Experiment : Phonological	-0.022	0.021	-1.037	0.300

Table C32. Results of the RT analysis for comprehension and production particiapants in the End-of-Session test in both Exp 1 and Exp 2, comparing semantic and phonological similarity conditions against the unrelated condition and their interaction with experiment.

Fixed Effect	Coefficient	SE	t	p-value
Intercept	8.541	0.028	307.861	< 0.001
Experiment	0.016	0.036	0.458	0.648
Semantic	0.095	0.020	4.760	< 0.001
Phonological	0.023	0.028	0.824	0.414
Experiment : Semantic	0.008	0.029	0.283	0.777
Experiment : Phonological	-0.022	0.027	-0.819	0.413

Footnotes

1. Note that R packages that allow pairwise comparisons (i.e., three comparisons for three levels of one factor) such as emmeans and multcomp currently do not handle interactions among multiple factors properly. Since one of our main questions is whether mode and contextual similarity interact, we ran two sets of models with planned comparisons that directly address the questions of interest. However, we do use emmeans in the combined analyses to both test the robustness of the effect of mode and to directly compare the effect sizes of production training against study and comprehension.
2. There were no differences in the pattern of results between Strict and Lenient coding for any analyses, therefore only results for Lenient coding are reported.
3. As in Experiment 1, there were no differences in results between Strict and Lenient coding for any analyses, therefore only results with Lenient coding are reported here.
4. The pattern of results is robust against adding random slopes for Similarity by Subject, Mode by Item, and Experiment by Item.